Contents

1. INTRODUCTION 3
OVERVIEW OF STUDY 3
ROUTE GUIDE FOR USERS 3
2. STUDY RELATED INFORMATION 4
DESIGN OVERVIEW 4
DATA COLLECTION AND RESPONSE OUTCOMES 5
OVERVIEW 5
FIGURE 1. TIMING OF DATA COLLECTION 5
DATA COLLECTION 5
PANEL MEMBERSHIP AND PANEL MAINTENANCE 7
RESPONSE OUTCOMES – WAVE 1 8
RESPONSE OUTCOMES – WAVE 2 8
DATA PROCESSING AND CLEANING 13
DOCUMENTATION OF THE QUESTIONNAIRES: MODULES AND QUESTIONS 13
READING THE QUESTIONNAIRES 14
SUMMARY OF QUESTIONNAIRE MODULES 15
CHANGES TO THE QUESTIONNAIRE 18
OTHER FIELDWORK MATERIALS 18
SAMPLE DESIGN 18
GENERAL POPULATION SAMPLE COMPONENT. 19
GENERAL POPULATION COMPARISON SAMPLE COMPONENT. 19
ETHNIC MINORITY BOOST SAMPLE. 20
FORMER BHPS SAMPLE 21
SAMPLE STATUS AND FOLLOWING RULES 21
SAMPLE DESIGN VARIABLES AND ANALYSIS 22
WEIGHTING ADJUSTMENTS FOR THE WAVE 2 RELEASE 23
SELECTING THE CORRECT WEIGHT FOR YOUR ANALYSIS 24
NAMING CONVENTIONS FOR WEIGHTING VARIABLES 27
TECHNICAL DETAILS 28
IMPUTATION OF INCOME VARIABLES 36
WHAT DO WE IMPUTE? 37
IMPUTATION PROCEDURES 38
ITEM NON-RESPONSE ON INCOME VARIABLES IN THE INDIVIDUAL QUESTIONNAIRE 38

ITEM NON-RESPONSE FOR INCOME VARIABLES IN THE PROXY QUESTIONNAIRE 40
INDIVIDUAL NON-RESPONDENTS WITH NO PROXY QUESTIONNAIRE 40
CODING 41
3. FILE AND VARIABLE INFORMATION 41
INFORMATION ABOUT THE BHPS SAMPLE COMPONENT 41
VARIABLE INFORMATION – OVERVIEW BASIC AND DERIVED VARIABLES 43
VARIABLE NAMING AND LABELLING CONVENTIONS 43
LEARNING ABOUT THE STUDY VARIABLES 43
IDENTIFIERS AND USEFUL VARIABLES 44
TABLE 9. SOME USEFUL VARIABLES 44
DOCUMENTATION OF DERIVED VARIABLES 45
EXAMPLE CODE FOR MATCHING FILES 46
EXAMPLE 1: DISTRIBUTING HOUSEHOLD LEVEL INFORMATION TO INDIVIDUAL LEVEL 46
EXAMPLE 2: SUMMARISING INDIVIDUAL LEVEL INFORMATION AT THE HOUSEHOLD LEVEL 47
EXAMPLE 3: MATCHING INDIVIDUALS WITHIN A HOUSEHOLD 48
EXAMPLE 4: USING THE EGOALT FILE TO CREATE HOUSEHOLD COMPOSITION VARIABLES 49
EXAMPLE 5: MERGING INDIVIDUAL FILES ACROSS WAVES INTO LONG FORMAT 49
EXAMPLE 6: MERGING INDIVIDUAL FILES ACROSS WAVES INTO WIDE FORMAT 50
PRESERVING CONFIDENTIALITY 51
PARADATA 51
4. DATA ACCESS 52
5. CITATIONS AND ACKNOWLEDGEMENTS 53
6. REFERENCES 53

Understanding Society: UK Household Longitudinal Study: Wave 1-2, 2009-2011 User Manual

1. Introduction

Overview of study

Understanding Society -- the UK Household Longitudinal Study (UKHLS) --is a longitudinal survey of the members of approximately 40,000 households in the United Kingdom (England, Scotland, Wales and Northern Ireland). Households recruited at the first round of data collection are visited one year later to collect information on changes to their household and individual circumstances. Interviews are carried out face-to-face in respondents' homes by trained interviewers. Wave 1 data collection took place between January 2009 and January 2011. Wave 2 took place between January 2010 and April 2012.

Understanding Society is funded by the Economic and Social Research Council and with funding from multiple government departments (the Department for Work and Pensions, the Department for Education, the Department for Transport, the Department for Culture, Media and Sport, the Department for Communities and Local Government, the Department of Health, the Scottish Government, the Welsh Assembly Government, the Northern Ireland Executive, the Department for Environment, Food and Rural Affairs, and the Food Standards Agency). The scientific leadership team is from the Institute for Social and Economic Research (ISER) of the University of Essex, the University of Warwick, and the Institute of Education, University of London. Professor Nick Buck is the principal investigator. Fieldwork is conducted by the National Centre for Social Research (NatCen) with collaboration with the Central Survey Unit of the Northern Ireland Statistics and Research Agency (NISRA) in Northern Ireland.

The overall purpose of *Understanding Society* is to provide high quality longitudinal data about subjects such as health, work, education, income, family, and social life to help understand the long term effects of social and economic change, as well as policy interventions designed to impact upon the general well-being of the UK population.

Route guide for users

This release has data for the General Population and the Ethnic Minority boost (EMB) sample. Former participants of the British Household Panel Survey (BHPS) are part of *Understanding Society* from Wave 2 (http://www.iser.essex.ac.uk/bhps). The BHPS is a household panel survey of around 8,000 households in the UK, which has completed 18 annual waves of data collection and has been run by ISER since it began in 1991. Data from the BHPS can be obtained from the UK Data Archive (SN5151 British Household Panel Study, Waves 1-18, 1991-2009, http://www.esds.ac.uk/findingData/snDescription.asp?sn=5151.) Data from the Innovation

Panel, a separate survey intended to support methodological research (www.understandingsociety.org/design/innovation/default.aspx). Data from the Innovation Panel has been released through the UK Data Archive (SN6849 *Understanding Society:* Innovation Panel, Waves 1-2, 2008-2009, http://www.esds.ac.uk/findingData/snDescription.asp?sn=6849).

The Ethnic Minority Boost sample was undertaken to produce enough cases to study households and individuals from five major ethnic groups in the UK. The boost sample receives an additional five minutes of questions related to content areas that may particularly involve them. The General Population Comparison sample component is also asked these questions.

As an introduction to the data and documentation we recommend the following reading:

1. The summary of the general questionnaire content (Section, 2, Documentation of the questionnaires: modules and questions, Reading the questionnaire) and notes on naming conventions (Section 3 Variable naming and labeling conventions

2. Study level information is in Section 2. This includes sections on sample design, weighting adjustments and data collection and response outcomes.

3. Variable level descriptions of the data can be found on the *Understanding Society* website (http://data.understandingsociety.org.uk/documentation). The online documentation has extensive links between questions and detailed views of variables and data files. There is also a search facility for searching questions, variables, modules, and datafiles.

4. The example Stata code for matching variables from different records (Section 3, Example code for matching files).

In assembling the documentation, we have drawn upon the documentation for the British Household Panel Survey (Taylor 2010) (see also http://www.iser.essex.ac.uk/bhps).

2. Study Related Information

Design overview

Understanding Society is a panel survey of households with yearly interviews. Data collection for a single wave is scheduled across 24 months. The study began with a representative probability sample of households. There is an extended discussion of sample design below and in (Lynn, 2009). Adult household members (age 16 or older) are asked questions and the same individuals are re-interviewed in successive years to see how things have changed. Household members aged 10-15 years are asked to complete a short self-completion youth questionnaire. Children become eligible for a full interview once they reach the age of 16.

The overall study has multiple sample components. In the mainstage survey there is the a) General Population Sample, with its subset the General Population Comparison Sample, b)

the Ethnic Minority Boost Sample, and c) participants from the British Household Panel Study. The instruments for the first three components are the same except the EMB sample and the General Population Comparison sample have an "Extra five minutes" of questions specifically relevant to ethnic minority communities e.g., ethnic identity and remittances.

In addition, there is a separate survey, the Innovation Panel (IP), which is fielded in the year before the mainstage survey. It tests varying measurement issues, and its instruments are somewhat different from the mainstage survey. The IP can be accessed through the UK Data Service, SN 6849.

Data collection and response outcomes

Overview

Figure 1 shows the timing of data collection for the data included in this release, and for the previous wave of data collection for each of the two samples in this release, namely wave 1 of *Understanding Society* and wave 18 of the BHPS. The BHPS may be accessed via SN 5151, from the UK Data Service.

Data collection for a single wave is scheduled across 24 months. There is some variation in that pattern. The data collection for Northern Ireland and the former BHPS sample component takes place in the first 12 months of the wave.

Most of the data collection is conducted face-to-face via computer aided personal interview (CAPI). There are also self-completion instruments for youth and adults.

Figure 1. Timing of Data Collection

2008				2009				2010				2011			
Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
		BHPS Wave 18													
				Wave 1 year 1				Wave 1, year 2							
								Wave 2, year 1				Wave 2, year 2			

Q - Quarter

Data Collection

The players, who does what

ISER, together with NatCen and the Central Survey Unit of Northern Ireland Statistics and Research Agency (NISRA), work closely together on all aspects of data collection, implementing an agreed set of survey procedures designed to ensure adequate response and effective data quality.

ISER has the primary responsibility for design work. NatCen manages fieldwork, editing, and coding and data-entry. It also advises on the design of all research instruments. NISRA collaborates with NatCen and is responsible for fieldwork in Northern Ireland. ISER plays a major role in quality control through specification of fieldwork practices; survey materials; editing and coding requirements; and inspecting and analysing weekly fieldwork progress reports. This working relationship is reinforced by an agreed set of survey-specific procedures to ensure adequate response and effective data quality. Full details of these, and other technical aspects of the data collection and fieldwork, coding, and data processing are

found in the *Technical Reports,* published each wave on the *Understanding Society* website (see http://data.understandingsociety.org.uk/.)

Getting ready for fieldwork

Prior to the first wave of the main *Understanding Society* survey, there were two small pilot studies and a dress rehearsal. A cognitive pilot of 70 individuals was conducted March – April 2008 to test screening and other questions relevant to the ethnicity strand. A translation pilot was conducted in June 2008: 50 interviews were carried out using Bengali and Punjabi translations of the questionnaire to see if there were problems with the operation of the translation program or problems with interviewing with the translated instruments. A run-through of all data collection instruments and procedures in 100 households, called a dress rehearsal, took place August-September 2008.

A pilot or run-in for Wave 2 tested all instruments and data collection procedures. For this wave, the data collection also focused on assessing any problems with integrating members of the former BHPS sample component, which includes a small segment conducted by telephone interviews. In all 237 households were issued. Of these, 91 were households interviewed in the Wave 1 pilot. The BHPS sample component was represented by households that were part of the BHPS between 1997 and 2001, the European Community Household Panel. Households for which we had a telephone number were issued to telephone interview to test the telephone interview instruments and procedures. The Wave 2 pilot took place September-October 2009.

Interviewers

Because of the demanding nature of *Understanding Society* we tried to use interviewers of above average levels of experience and ability. In Northern Ireland, the majority of interviewers had worked on the Northern Ireland component of the BHPS (the Northern Ireland Household Panel Survey), and were familiar with the design and operation of *Understanding Society*.

In addition to general interviewer training, interviewers working on *Understanding Society* attended a survey-specific face-to-face briefing. Generally around 12-20 interviewers attended each briefing, along with two or three briefing managers or area managers. The briefings were led by at least one researcher from NatCen with the majority also attended by ISER staff. The briefings in Wave 1 took place across the UK; Belfast, Birmingham, Brentwood, Bristol, Derby, Edinburgh, Glasgow, Leeds, London and Manchester. Similar topics and and locations were used for the Wave 2 briefings.

These one-day briefings had morning sessions devoted to fieldwork procedures, including dealing with the administrative forms to record contact information, and how to deal with the complexities of multiple dwelling units and multiple households. The afternoon was spent discussing the survey content and reviewing and working with the Blaise computer aided personal interview (CAPI) instrument.

Interviewers were assigned to specific areas. For Wave One, 911 interviewers were employed to cover 3,517 areas in the sample. The number of interviewers briefed in Wave 2 was 819.

Fieldwork

Fieldwork for Wave 1 was somewhat different because we did not know who was in the sample. Before contacting any of their sample in Wave 1, interviewers mailed an introductory card from ISER to all sampled addresses (addressed to "The Occupier"), together with a small leaflet outlining the purpose of the survey. The interviewer called within a week of the mailing. At the end of the first interview, all participating households received a more detailed brochure, giving further information about the survey and thanking respondents for participating.

A minimum of six calls is made at each sampled address before it is considered a non-contact. Interviewers are encouraged to make further calls, if possible. If there was a potential for success, a special conversion letter is sent to households which had refused to participate or had not been contacted.Post interview quality control is carried out with a telephone recall on 10% of all completed interviews.

Interviewers upload their work daily, including information about all the calls they have made, whether or not there was any response. This information is collated by NatCen to construct a weekly field progress monitor report for ISER.

Panel membership and panel maintenance

The rules for following individual respondents over time stem from the composition of the household. Individuals found at selected households in the first wave were designated as Original Sample Members (OSM). We attempt to maintain OSM respondents as part of the sample as long as they live in the UK. In addition, births to an OSM mother are also classified as OSM. Individuals joining the household of an OSM after enumeration of the household at Wave 1 are Temporary Sample Members (TSM). One deviation from this is for individuals who were not an ethnic minority within the households selected as the ethnic minority boost sample. At Wave 1, these individuals were classified as TSMs. We attempt to interview TSM participants in successive waves as long as they live in the household of an OSM.

A male TSM who fathers a child with an OSM female becomes a Permanent Sample Member (PSM). PSMs are treated in the same way as OSMs in the following rules. In sum, TSMs are not followed for interviews when they leave the household, but OSMs and PSMs are. An exception to these sample status rules is that at Wave 1 individuals who were not of an ethnic minority within an ethnic minority boost household were classified as TSMs.

For panel maintenance, ISER maintains a database of information on respondents so we can send communications to them and to allocate interviewers. This information is vital for minimising attrition. The data base builds on contact information collected during the survey interviews, and is updated throughout the year. There are, for example, new addresses, household splits and moves out of the country or into an institution. Change of address cards were also returned to ISER in cases where a whole household moved or a new resident returned the card giving the forwarding address. Finally, it is possible for ISER to be notified of some deaths through this means.

A between wave mailing is also used to help maintain contact with participants and update addresses. The mailing has a report of research findings, an address confirmation slip and

materials to encourage registration with the Participant website. The participant website can be seen at http://participants.understandingsociety.org.uk/.

Response outcomes – Wave 1

The Wave One mainstage fieldwork started on 8th January 2009 and ended on the 7th March 2011 (including the re-issue period). In total, interviews were achieved in 30,169 households (26,089 in the General Population Sample, 4,080 in the ethnic minority boost sample), with full or proxy interviews with 50,994 individuals (43,674 in the General Population Sample and 7,320 in the ethnic minority boost sample).

Tables 1 and 2 below present the household and individual response rates for Wave 1. The individual response rates are for co-operating households only.

Table 1. Household response rates among eligible households, Wave 1

	General Population Sample			Ethnic Minority Boost
	Great Britain	Northern Ireland	Total	
Responding	57.1%	60.9%	57.3%	39.9%
Non-contact	8.1%	11.0%	8.3%	28.0%
Refusal	33.9%	27.4%	33.6%	29.0%
Other	0.8%	0.7%	0.8%	3.1%
N	43,267	2,107	45,374	10,077

The response rates for the ethnicity boost sample component do not make any correction for the probability of non-interviewed cases being ineligible. The estimated response rate taking this factor into account is substantially higher.

Table 2. Individual response rates, Wave 1

	General Population Sample			
	Great Britain	Northern Ireland	Total	Ethnic Minority Boost
Full interview	82.0%	77.3%	81.8%	72.4%
Proxy interview	5.3%	3.5%	5.2%	6.9%
Refusal	6.5%	9.2%	6.7%	8.7%
Other non-interview	6.1%	9.9%	6.3%	12.1%
N	47,615	2,584	50,199	9,237

Response outcomes – Wave 2

Table 3, below, shows the household response rates for Wave 2 of the UKHLS. The table separates the different samples. The General Population Sample (GPS) consists of respondents in Great Britain and Northern Ireland. The ethnic minority boost (EMB) households are only located in Great Britain. The former-BHPS sample consists of the *Living in Britain* sample (started in 1991), the *Living in Scotland* and *Living in Wales* boost samples (started in 1999) and the *Northern Ireland Household Panel Survey* (NIHPS, started in 2001), also a boost sample.

Ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals or emigrated from the UK. For the former-BHPS samples, ineligible households would also include households which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

Responding households are those in which the household is successfully enumerated, the household questionnaire is completed and all eligible adults give an individual interview. Partially responding households are those where the household is enumerated and a household questionnaire is done, and at least one eligible adult but not all eligible adults complete an individual interview.

Household response rates were higher in Northern Ireland than in the rest of the UK. The household response rate for the continuing *Understanding Society* General Population Sample was 76.8% in Great Britain, and 81.9% in Northern Ireland. The household response rates for the former-BHPS samples were similar to the *Understanding Society G*eneral Population Sample. Among the samples in Great Britain, the *Living in Britain* households had the highest response rate at 77.2%. The *Living in Wales* households had a similar response rate to the *Living in Britain* sample (76.8%) whilst *Living in Scotland* had a lower response rate at 73.5%. The NIHPS had a higher household response rate than the Great Britain samples, with 84.8%. The response rates for the BHPS samples in Great Britain were disappointing, given that this was, in effect, Wave 19 for many households. However, the lower response rate may have been due to the change in the fieldwork agency, in interviewers, in the survey name and in the logo. Interestingly, in Northern Ireland where the survey name and logo changed, but the fieldwork agency and so the interviewer stayed the same as in NIHPS, the response rate was much higher.

Table 3: Household response rates, Wave 2*

	UKHLS GP sample		**EMB**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	16,003 61.8%	873 65.3%	2,030 49.3%	3,112 66.5%	793 64.3%	833 64.1%	990 73.4%	24,634 61.7%
Partially responding	3,888 15.0%	221 16.5%	749 18.2%	504 10.8%	114 9.2%	165 12.7%	153 11.4%	5,794 14.5%
All responding	19,891 76.8%	1,094 81.9%	2,779 67.5%	3,616 77.2%	907 73.5%	998 76.8%	1,143 84.8%	30,428 76.2%
Non-contact	1,116 4.3%	22 1.7%	299 7.3%	217 4.6%	73 5.9%	62 4.8%	33 2.5%	1,825 4.6%
Untraced mover	1,450 5.6%	50 3.7%	411 10.0%	181 3.9%	49 4.0%	50 3.9%	43 3.2%	2,235 5.6%
Refusal	3,359 13.0%	162 12.1%	600 14.6%	648 13.8%	199 16.1%	185 14.2%	117 8.7%	5,281 13.2%
Other non-interview	94 0.4%	8 0.6%	28 0.7%	20 0.4%	6 0.5%	5 0.4%	12 0.9%	173 0.4%
Total*	25,910	1,336	4,117	4,682	1,234	1,300	1,348	39,942

* Base is all households issued to the field for wave 2, minus any found to have become ineligible.

Non-contact rates were lower in Northern Ireland than in Great Britain. The level of untraced movers was higher for the *Understanding Society G*eneral Population Sample in Great

Britain than in the former-BHPS samples. The level of non-contact and untraced movers were highest in the ethnic minority boost samples, possibly reflecting the younger average age of this sample, the concentration in large urban areas and the higher level of mobility. Within the former-BHPS samples, the level of untraced movers was higher than in the past. This is likely to be due to the increased gap between waves of interview. The interviews for the former BHPS sample for Wave 2 of the UKHLS took place throughout 2010 and into the early months of 2011. The previous interview for most of these households was between September and December 2008. As the gap between the Wave 18 BHPS interview and the Wave 2 *Understanding Society* interview increased, so did the level of untraced movers.

Refusals, as well, were generally higher in Great Britain than in Northern Ireland. Refusals are expected to be higher at the second wave of a longitudinal study than at subsequent waves. The higher than expected refusal rate for the former BHPS sample, particularly those in Great Britain, may be due to the aforementioned change in the name and logo of the study as well as the change in fieldwork agency and thus, for most households, a change of interviewer.

Table 4, below, shows the cross-sectional response rates for adults in Wave 2. Where a household responded, we have an individual-level outcome for all adults. Where a household did not respond, we have assigned the household non-response outcome to the adults who were issued to that household. From this we can see, for example, that we were not able to interview 7,229 adults in the UKHLS General Population Sample in Great Britain because they were residing in households who refused to participate at Wave 2. In the Great Britain samples of the former BHPS there is a relatively small group of households who only give telephone interviews.

On a longitudinal study, such as the UKHLS, researchers are typically interested in having pairs of observations on the same individual to investigate individual-level change over time. Table 5, below, takes as the baseline all those who gave a full interview at the previous wave, and shows their outcome at Wave 2. For the former BHPS sample, the previous wave was Wave 18 (LIB), Wave 10 (LIS and LIW) or Wave 8 (NIHPS), all collected in 2008. Once more, we see that there is a higher re-interview rate in the Northern Ireland samples than in Great Britain. The lowest re-interview rate is in the ethnic minority boost sample, largely due to a higher level of non-contacted households or households who moved but could not be traced. Interestingly, the re-interview rate was higher in the General Population GB sample than in the three samples that made up the former-BHPS GB samples. Overall, in the Waves 1 and 2 data, pairs of observations are available for 45,836 adults. If proxy and telephone interviews are included, this increases to 47,282 adults.

For more detail, please see the working paper on non-response and attrition (Lynn et al., 2012).

Table 4: Cross-sectional individual adult response rates by sample origin

	UKHLS GP sample		**EMB**	**Former-British Household Panel Survey**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	32,381 60.8%	1,770 62.3%	4,978 46.3%	6,140 61.6%	1,461 58.1%	1,651 59.6%	2,008 71.7%	50,389 59.4%
Proxy interview	2,722 5.1%	87 3.1%	615 5.7%	253 2.5%	49 2.0%	86 3.1%	58 2.1%	3,870 4.6%
Telephone interview	--	--	--	202 2.0%	66 2.6%	58 2.1%	--	326 0.4%
Other non-interview	1,184 2.2%	126 4.4%	472 4.4%	200 2.0%	58 2.3%	64 2.3%	107 3.8%	2,211 2.6%
Refusal	2,104 4.0%	218 7.7%	511 4.8%	341 3.4%	92 3.7%	114 4.1%	133 4.8%	3,513 4.1%
Household non-contact	3,338 6.3%	125 4.4%	1,156 10.8%	555 5.6%	210 8.3%	155 5.6%	111 4.0%	5,650 6.7%
Household refusal	7,229 13.6%	350 12.3%	1,743 16.2%	1,493 15.0%	400 15.9%	427 15.4%	203 7.2%	11,845 14.0%
Household other non-interview	118 0.2%	5 0.2%	60 0.6%	29 0.3%	9 0.4%	6 0.2%	4 0.1%	231 0.3%
Household untraced	4,178 7.9%	159 5.6%	1,207 11.2%	754 7.6%	172 6.8%	208 7.5%	178 6.4%	6,856 8.1%
Total	53,254	2,840	10,742	9,967	2,517	2,769	2,802	84,891

Table 5: Longitudinal individual re-interview rates (adults) by sample origin: Full interview at the previous wave

	UKHLS GP sample		EMB	Former-British Household Panel Survey				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	29,646 74.3%	1,640 81.0%	4,200 62.2%	5,633 69.4%	1,335 67.8%	1,507 67.6%	1,875 83.3%	45,836 72.4%
Proxy interview	775 1.9%	16 0.8%	188 2.8%	97 1.2%	17 0.9%	38 1.7%	15 0.7%	1,146 1.8%
Telephone interview	--	--	--	184 2.3%	59 3.0%	57 2.6%	--	300 0.5%
Other non-interview	334 0.8%	22 1.1%	157 2.3%	73 0.9%	16 0.8%	28 1.3%	32 1.4%	662 1.1%
Refusal	316 0.8%	25 1.2%	94 1.4%	53 0.7%	13 0.7%	15 0.7%	11 0.5%	527 0.8%
Household non-contact	1,890 4.7%	68 3.4%	500 7.4%	376 4.6%	126 6.4%	96 4.3%	34 1.5%	3,092 4.9%
Household refusal	4,144 10.4%	167 8.3%	734 10.9%	965 11.9%	245 12.4%	260 11.7%	109 4.8%	6,633 10.5%
Household other non-interview	65 0.2%	2 0.1%	31 0.5%	18 0.2%	3 0.2%	2 0.1%	2 0.1%	123 0.2%
Household untraced	2,252 5.6%	63 3.1%	639 9.5%	439 5.4%	105 5.3%	140 6.3%	105 4.7%	3,744 5.9%
Household ineligible	507 1.3%	22 1.1%	208 3.1%	280 3.5%	51 2.6%	86 3.9%	68 3.0%	1,222 1.9%
Total	39,929	2,025	6,751	8,118	1,970	2,229	2,251	63,285

Data Processing and cleaning

NatCen delivers the data for a sample month to ISER in batches. Delivery is scheduled for 4 months following the beginning of the fieldwork process, to allow time for interview re-issue, coding, and data entry from paper documents, e.g., the self-completion instruments.

Data is delivered as SPSS system files, which are then exported to triple-S data exchange format and imported into a SIR database.

Quality control processes include extensive data checking to ensure that the data conform to the expected structure and to the routing and range constraints defined by the questionnaire specifications.

Data anomalies are investigated to determine whether they are related to:
1) the invalid specification of the questionnaire; 2) the incorrect scripting of the questionnaire; 3) a failure to specify that a particular constraint should be included in the questionnaire; 4) an incorrect implementation of the check, or; 5) a problem in exporting and/or delivering the data. After investigation, steps may include correcting the specification, data editing, reporting the error to NatCen to be fixed in a subsequent delivery and/or a quality feedback report suggesting changes to the questionnaire or field practice in subsequent waves.

Batch-specific databases are merged into a single database, from which anonymised data is exported for the creation of public use files. Data distributions are also checked for theoretical and statistical plausibility. This checking is done through direct scrutiny and by analyses which 'road-test' the data.

Documentation of the questionnaires: modules and questions

The text of the questionnaires in pdf format is part of the documentation provided through the UK Data Archive. Questionnaires can also be found: http://data.understandingsociety.org.uk/documentation/mainstage/questionnaires. The documentation is for the mainstage survey (household and individual) and the adult and youth self-completion instruments. The instruments are an important source of information about the wording of individual questions, who was asked, and what questions precede and follow.

Most of the interview is conducted with a computer-assisted personal interview (CAPI). The CAPI instrument governs the flow of questions and recording of answers, but it is not convenient for documentation. On the study website, we are presenting the questionnaire in PDF format. Similar to other PDF documents, the text of the questionnaire can be searched for specific words, such as variable names or words in questions. They are also convenient for printing sections of the instrument. The self-completion instruments are also displayed in pdf to correspond to the way they appeared to participants (except they have been annotated with variable names).

The principal adult questionnaires are organized in modules. Modules can be searched for in the online documentation system. In the pdf-formatted questionnaire, clicking on entries in the table of contents will advance you to the beginning of that module.

Instruments and survey materials were translated into multiple languages: Bengali; Punjabi in Urdu and Gurmukhi scripts; Welsh; Arabic; Somali; Cantonese; Urdu and Gujarati. Translated documents can be requested by email from info@understandingsociety.org.uk.

Reading the questionnaires

Figure 1 shows a marked up sample page, providing information for how to interpret the questionnaire text. Note that the variable names in the questionnaire do not have the wave prefix (a_).

Figure 1. Mark-up of household questionnaire

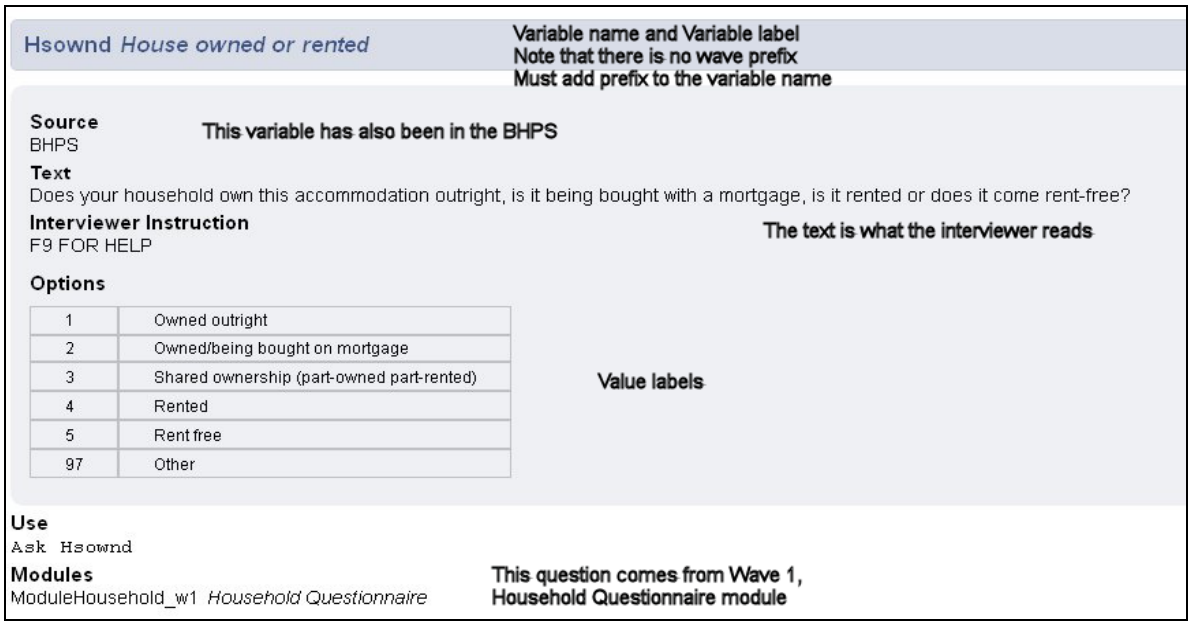
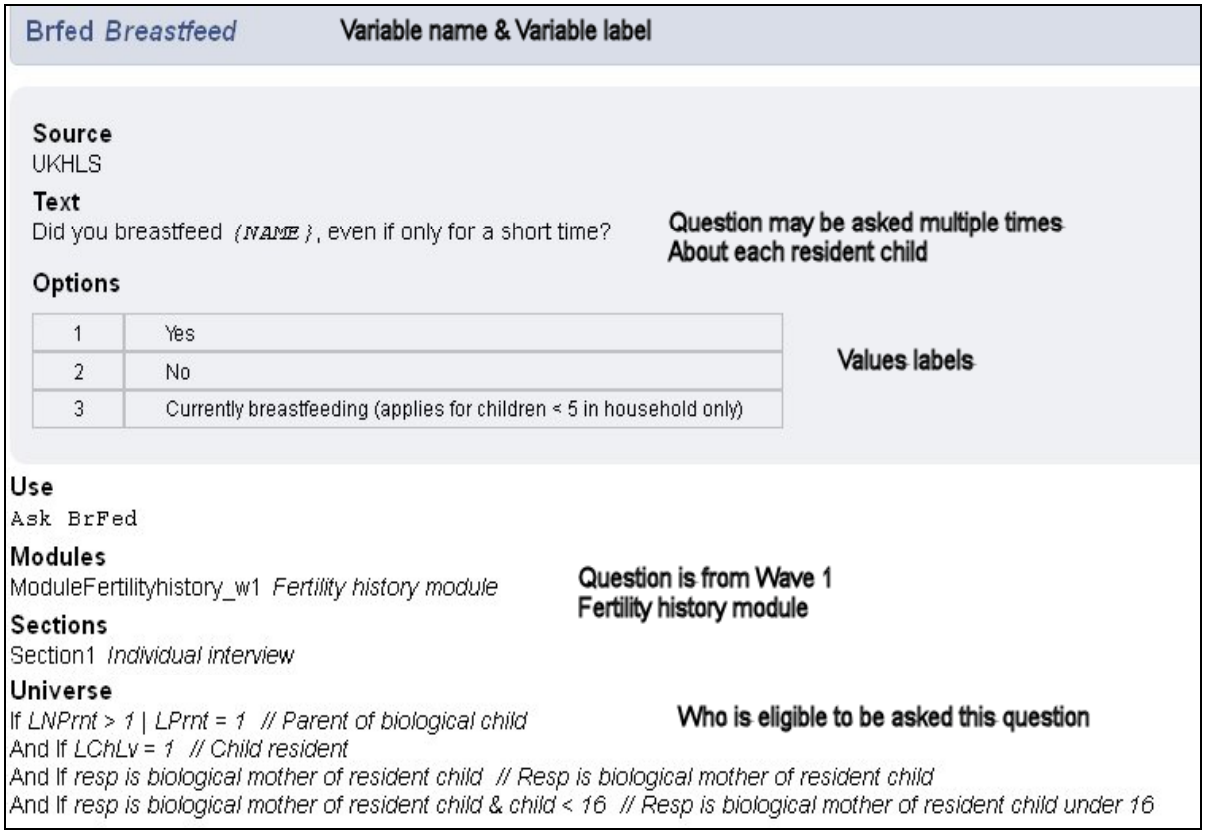

Hsownd House owned or rented
Variable name and Variable label
Note that there is no wave prefix
Must add prefix to the variable name
Source
BHPS
This variable has also been in the BHPS
Text
Does your household own this accommodation outright, is it being bought with a mortgage, is it rented or does it come rent-free?
Interviewer Instruction
F9 FOR HELP
The text is what the interviewer reads
Options
1 Owned outright
2 Owned/being bought on mortgage
3 Shared ownership (part-owned part-rented)
4 Rented
5 Rent free
97 Other
Value labels
Use
Ask Hsownd
Modules
ModuleHousehold_w1 Household Questionnaire
This question comes from Wave 1,
Household Questionnaire module

Figure 2 shows a marked up sample page from the individual interview. The question is more complex. The question is asked about each natural or biological child, so multiple variables are associated with the question for each natural child. The variables are located in the data file A_NATCHILD, which has one record for each natural child.

Figure 2. Mark-up of question with looping from individual questionnaire


Brfed Breastfeed
Variable name & Variable label
Source
UKHLS
Text
Did you breastfeed (NAME), even if only for a short time?
Question may be asked multiple times
About each resident child
Options
1 Yes
2 No
3 Currently breastfeeding (applies for children < 5 in household only)
Values labels
Use
Ask BrFed
Modules
ModuleFertilityhistory_w1 Fertility history module
Question is from Wave 1
Fertility history module
Sections
Section1 Individual interview
Universe
If LNPrnt > 1 | LPrnt = 1 // Parent of biological child
Who is eligible to be asked this question
And If LChLv = 1 // Child resident
And If resp is biological mother of resident child // Resp is biological mother of resident child
And If resp is biological mother of resident child & child < 16 // Resp is biological mother of resident child under 16

Summary of questionnaire modules

About half of the questionnaire content is collected annually, with additional modules collected different intervals, often every two to three years. The long-term content plan summarizes the pattern that has been collected or planned: http://www.understandingsociety.org.uk/design/content/outlines.aspx.

Table 6 Summary of Questionnaire Modules in Waves 2 and 1

Module	Asked of a subset in Wave 2?	Wave 2	Also in Wave 1
Demographics	new entrants	X	X, of all
Initial Conditions	new entrants	X	X, of all
Own First Job*	new entrants, excluding rising 16's	X	X, part of employment status history asked in months 1-6
Parental Education	if not asked in Wave 1	X	X
Educational Aspirations	asked of full-time students	X	
Young adults	age 16-21	X	
Family background	new entrants	X	
Ethnicity and National Identity	new entrant	X	X, of all
Childhood Language		X	
Ethnic Identity	EMB, GPC, LDA	X	

Religion	new entrant and EMB, GPC, LDA	X	X, of all
General Health		X	X (see Health and Disability module)
Nutrition		X	
Physical Activity		X	
Smoking History		X	
Disability		X	X (see Health and Disability module)
Caring		X	X
Partnership History	new entrants	X	X, of all ever in partnership
Fertility History	new entrants and ever in partnership	X	X, has adopted or biological children
Annual Event History	interviewed in last wave	X	
Current Employment		X	X
Employees		X	X
Self-employment		X	X
Commuting Behaviour	employee or self-employed not in home	X	
Job Satisfaction	employed	X	X
Physical Work	in paid work	X	
Work Conditions	employee	X	
Non-employment	no paid work, no job	X	X
Second jobs		X	X
Voluntary Work		X	
Charitable Giving		X	
Childcare	Responsible for children	X	X
Unearned Income and State Benefits		X	X
Household Finances		X	X
Personal Pensions		X	
Savings		X	
Retirement Planning	age 45, 50, 55, 60, 65 and not retired	X	
Domestic Division of Labour	married, civil partnership, cohabiting and live with partner	X	One item on hours of housework asked in first 6 months at w1
Politics		X	X
Political Engagement	EMB, GPC, LDA comparison	X	
General Election	interviewed May-December 2010	X	
Leisure, Culture and Sports		X	
Leisure Access		X	
Positive and Negative Events		X	
Interviewer		X	X

Observations			
Proxy	asked about those not able to be interviewed in person	X	X
In Wave 1			
Language		but see Childhood language	X
Migration History	EMB, GPC,LDA		X
Employment Status History			X (asked of first 6 months of first year)
Discrimination			X , EMB, GPC, LDA and job variables
Parents and Children			X, responsible mother and responsible father of children
Family Networks			X
Remittances			X, EMB, GPC, LDA
Harassment			X, EMB, GPC, LDA
Environmental Behaviour			X
Consents for linkage to health and education administrative records			X
EMB – ethnic minority boost; GPC – General Population Comparison, LDA – Low density area			
* Own first job will be asked of OSMs and new entrants in Wave 4			

The self-completion questionnaires are not divided into modules, but Table 7 summarizes the content in waves 2 and 1.

Table 7. Summary of Adult Self-Completion Questionnaires in Waves 2 and 1

	Wave 2	Wave 1
GHQ-12	X	X
Satisfaction: life, and other	X	X
Alcohol consumption	X	
Control	X	
Positive and negative social support	X	
SF-12	X	in adult interview Health and Disability module
Gender role opinions	X	
Identity	X	
Sleep		X
Environmental attitudes and		X

beliefs		
Neighbourhood belonging and participation		X
Trust		X
Short Warwick-Edinburgh Mental Well Being Scale		X
Attitudes to risk		X
Partnership relationship: quality, activities, happiness		X

The content of the Youth self-completion instruments is summarized in pages 6-7 of the long-term content plan:
http://www.understandingsociety.org.uk/design/content/outlines.aspx.

Changes to the questionnaire
Questionnaire changes have been made under certain circumstances. At the end of the first six months of data collection in Wave 1, multiple variables were dropped because of the length of the interview, e.g., cutting of the employment history module. At the same time other modifications were made, e.g, in question format. Notes about these changes have been documented in the variable view of the online documentation system.

Other fieldwork materials
Other fieldwork materials are also on the website: http://data.understandingsociety.org.uk/documentation/mainstage-fieldwork. One example is the Showcards, which are used to help respondents with their answers. Showcards are referenced in the questionnaire. Project Instructions were prepared for interviewer training and to serve as a resource in data collection. Documents for communicating with participants are also included on the website. In Wave 1, we asked for consent to link to administrative health and education records. The information leaflets and consent forms are in this section of the study website.

The Address Record Form (ARF) is an important source of information about responding and non-responding households. It has the call record, observations on characteristics of accommodation and households, and household outcomes. In Wave 1 there are several different versions of the ARF. The first distinction is between the General Population Sample (GP) and the Ethnic Minority Boost Sample (EB). The versions labeled ARF are longer because they include questions for screening household members for eligibility. ARF's labeled 2 or 3 are for addresses with multiple households and/or dwelling units. Finally, there are versions for ARF EB1 Year 1 or Year 2. This change in form was required by the change in selection criteria implemented in Year 2 of Wave 1 (see Berthoud et al., 2009 for more detail). The ARF screening card was a show card used during the screening interviews. Additional information about completion of the ARF can be found in the Project Instructions for Interviewers.
(http://data.understandingsociety.org.uk/documentation/mainstage/fieldwork-docments).

Sample design

The *Understanding Society* sample consists of a new large General Population Sample plus four other components (the Ethnic Minority Boost Sample, the General Population Comparison sample, the ex-BHPS sample and the Innovation Panel sample). The design of

all five components is described in more detail in an *Understanding Society* working paper (see Lynn 2009). The Innovation Panel is prepared as a separate study, which can be accessed via the UK Data Service, SN 6849. The General Population Sample is based upon two separate samples of residential addresses, one for England, Scotland and Wales and one for Northern Ireland. The England, Scotland and Wales sample is a proportionately stratified (equal probability), clustered sample of addresses selected from the Postcode Address File. The Northern Ireland sample is an unclustered systematic random sample of addresses selected from the Land and Property Services Agency list of domestic addresses.

General Population Sample component.
The sample for England, Scotland and Wales was selected in two stages. The first stage was to select a sample of postcode sectors to serve as primary sampling units. The second stage was to select addresses within each sampled sector. Prior to selection, any postcode sector with fewer than 500 residential addresses was first grouped with an adjacent sector and thereafter treated as a single sector. The list of all sectors was then sorted into twelve geographical strata, consisting of ten regions in England plus Scotland and Wales as separate strata. Within each of the twelve strata, sectors were sorted into three sub-strata based upon the proportion of household reference persons classified as non-manual workers, based on 2001 Census data. Within each of the 36 sub-strata, sectors were then sorted into three further sub-divisions based on population density (households per hectare) and within each of the 108 resultant sub-divisions, sectors were listed in order of ethnic minority density. From the sorted list, a systematic random sample of 2,640 sectors was selected, with probability proportional to the number of residential addresses in the sector. These sectors were then allocated systematically to 24 monthly samples, with 110 sectors in each monthly sample. Within each postal sector, 18 addresses were selected using systematic random sampling. The England, Scotland and Wales sample in this data release is therefore based upon an initial sample of 47,520 addresses.

In Northern Ireland, 2,395 addresses were selected in a single stage from the list of domestic addresses. In combination, this data release is therefore based upon a total of 49,915 addresses.

At each address, the final stage of sampling was carried out by field interviewers. This consisted of identifying persons to be defined as sample members. All persons resident at each sample address at the time the interviewer made contact were deemed to be a sample member, with the exception of the small proportion of addresses that contained more than three dwellings or households. In those cases, three dwellings or households were sub-sampled at random.

General Population Comparison Sample component.
The General Population Comparison Sample (GPCS) has one sampled address for 40% of the selected postal sectors in General Population Sample (GPS) component for Great Britain. In other words, of the 2,640 general population sectors, 60% of them (1,584) contain 18 GPS addresses and the other 40% contain 17 GPS addresses and one GPCS address. The persons in these households will be designated as members of the General Population Comparison sample, regardless of ethnic group membership. Members of the General Population Comparison sample are a random subsample of the General Population Sample component and they should be included in analyses of the General Population Sample component.

Ethnic Minority Boost Sample.
The Ethnic Minority Boost Sample was designed to provide at least 1,000 adults from each of five groups: Indian, Pakistan, Bangladesh, Caribbeans, and Africans.

The initial step was identifying postal sectors with relatively high proportions of relevant ethnic minority groups, based upon 2001 Census data and more recent Annual Population Survey data. The set of 3,145 sectors constituted approximately 35% of the sectors in Great Britain and covered between 82% and 93% of the population of the five ethnic minority groups.

The 3,145 sectors were sorted into four strata based on the expected number of ethnic minority households that would be identified by the sampling and screening procedures (see Berthoud et al., 2009 for details). All sectors were included for the stratum where a yield of three or more households was expected. In the other three strata, sectors were sub-sampled at rates of 1 in 4, 1 in 8, or 1 in 16 respectively. This was done to constrain the number of sectors that might have just one or two eligible sample households (or even none). The total number of postal sectors selected for inclusion in the ethnic minority boost sample was 771. Of these 6 were in Scotland, 7 were in Wales, and the remaining 758 were in England, with a concentration in London (412 sectors).

The number of addresses selected per postal sector ranged from 15 to 103. Sampling fractions varied across the sectors in a way designed to deliver target numbers of respondents in each target ethnic minority group with adequate statistical efficiency (see Berthoud et al., 2009 for more details). In sectors selected for both the General Population Sample component and the Ethnic Minority Boost sample, a single systematic sample of the required total number of addresses was selected and allocated in a systematic way to the two sample components, thus ensuring that both sample components are spread throughout the whole sector.

The final stage of sampling was done by the interviewers for the Ethnic Minority Boost sample, though its procedures were somewhat more complex. You can see the steps described in the Project Instructions for Interviewers (http://data.understandingsociety.org.uk/assets/476). At addresses containing more than three dwellings or households, the procedures to sub-select dwellings or households were as described above for the General Population Sample component. Within each household, rather than all resident persons becoming sample members, there were three additional steps:

- A “screen” was carried out to identify whether there were any persons from target ethnic groups in the household.
- A random mechanism was applied to certain target groups identified by the screen in order to select only a desired proportion into the sample (non-mixed Indian, African, Far Eastern, Middle Eastern). For other target groups, all resident persons were included in the sample (mixed Indian, Bangladeshi, mixed Caribbean, Sri Lankan, Chinese, Turkish).
- In households included in the sample in the previous two steps, all members of target ethnic groups were deemed to be members of the Ethnic Minority Boost sample (including children). All persons of other ethnic groups are not Ethnic Minority Boost

sample members. They will be interviewed as temporary sample members for so long as they remain co-resident with at least one Ethnic Minority Boost sample member.

The overall sampling fractions combine a) the probability of sampling the sector, b) the fraction of addresses selected within the sector, and c) the probability of a household being retained following the application of the random selection mechanism described above.

Former BHPS sample

The sample issued at Wave two consisted of all members from the BHPS sample who were still active at Wave 18 of the BHPS and who had not refused consent to be issued as part of the *Understanding Society* sample. It should be noted that the BHPS sample contains different components, including the original sample, where households were first selected in1991, boost samples in Scotland and Wales, first selected in 1999, and a Northern Ireland sample, selected in 2001. For further details of the BHPS sample, see section IV of the BHPS User Guide (http://www.iser.essex.ac.uk/bhps/documentation/vola/vola.html).

Sample status and following rules

There are three possible sample statuses: Original Sample Members (OSMs), Temporary Sample Members (TSMs), and Permanent Sample members (PSMs). The definitions are as follows.

Original Sample Members (OSMs).

All members of *Understanding Society* General Population Sample households enumerated at Wave 1, including absent household members and those living in institutions who would otherwise be resident, are Original Sample Members (OSMs). All ethnic minority members of an enumerated household eligible for inclusion in the Ethnic Minority Boost sample are OSMs. In the Innovation Panel, all members of households enumerated at Wave 1 and refreshment sample households enumerated at Wave 4 are OSMs. In all of these samples, any child born to an OSM mother after Wave 1 and observed to be co-resident with the mother at the survey wave following the child's birth is an OSM. In the formerBHPS sample, OSMs are those who were enumerated at the first wave of the sample from which they come (Wave 1 for the original sample, Wave 9 for the Scotland and Wales boost samples, Wave 11 for Northern Ireland) or who were subsequently born to an OSM mother or father (or both). From Wave 2 onwards of *Understanding Society*, in the former BHPS sample, as for the rest of the *Understanding Society* sample, only children born to an OSM mother will themselves become an OSM. OSMs, of all ages, are followed for interview and remain eligible as long as they are resident within the UK. They remain potentially eligible sample members for the life of survey.

The case may arise where the only OSM in the household is a child. Other household members are then TSMs so long as they are co-resident with the child, and therefore eligible for interview, even if the child is not yet old enough to be eligible for interview. If the OSM child moves house, they are followed to their new address and those living with the OSM child are eligible for interview. If the OSM child moves into an institution, where normally just the OSM/PSM would be interviewed and not co-residents, a split-off household is created containing only the OSM child and the household enumeration grid completed. The child OSM is an eligible sample member, even if they are not eligible for interview because of their age.

Temporary Sample Members (TSMs).

Any members of an enumerated household eligible for inclusion in the Ethnic Minority Boost sample at Wave 1 who are not from a qualifying ethnic minority are Temporary Sample Members (TSMs) at Wave 1. This was the only category of TSM at Wave 1. In all parts of the sample, any new person found to be co-resident in an OSM or PSM household after wave 1 is a TSM. This would include any child born to an OSM father after wave 1 but not an OSM mother and observed to be co-resident with the father (or any other OSM) at the survey wave following the child's birth. TSMs remain eligible for interview as long as co-resident in an OSM/PSM household. TSMs who are not co-resident in an OSM/PSM household are not followed and become ineligible for interview. TSMs are identified as re-joiners if they are subsequently found in an OSM/PSM household and then become eligible for interview.

Permanent Sample Members (PSMs).

PSMs are TSMs who are followed for interview after they no longer live with an OSM. This is done for substantive research reasons because of the additional contextual information they may provide for the analysis of OSMs. At present, there is only one category of PSM, but others may be defined in the future. Any TSM father of an OSM child born after Wave 1 and observed to be co-resident with the child at the survey wave following the child's birth is a PSM. PSMs remain potentially eligible for interview for the life of survey.

Sample design variables and analysis

As the sample design involves stratification, clustering and weighting, these design features affect standard errors and should therefore be taken into account in analysis. Appropriate variables are provided to allow the analyst to do this. The weighting variables are described in a separate section. Here we describe the stratification and clustering variables.

*w*_psu This is an indicator of the primary sampling unit (PSU) to which the sample member belongs. The prefix w_ denotes waves in general.

The value of *w*_psu does not change between waves, but for new sample entrants it is only defined from the wave at which they enter the sample. *w*_psu takes values in the following ranges:

1 – 575	former BHPS sample in England, Scotland and Wales	Identical to the BHPS variable *w*psu
701 – 1849	former BHPS Northern Ireland sample	Corresponds to initial (BHPS wave 11) sampled households, as these were selected in a one-stage design
2001 – 4640	UKHLS-GPS in England, Scotland and Wales	Corresponds to the postal sectors used as PSUs (see Lynn, 2009)
4644 – 7035	UKHLS-GPS in Northern Ireland	Corresponds to wave 1 sampled households, as these were selected in a one-stage design
7078 – 51776	UKHLS-EMB	Corresponds to wave 1 sampled households, as these were selected in a one-stage design within the high minority density domain (see Berthoud et al, 2009)

*w*_ strata	This indicates the sampling stratum from which the sample member was selected.

The value of *w*_strata does not change between waves, but for new sample entrants it is only defined from the wave at which they enter the sample. *w*_strata takes values in the following ranges or the value 701 for the ex-BHPS Northern Ireland sample

1 – 151	ex-BHPS sample in England, Scotland and Wales	Identical to the BHPS variable *w*strata
701	ex-BHPS Northern Ireland sample	Northern Ireland treated as a single stratum
2001 – 3320	UKHLS-GPS in England, Scotland and Wales	Corresponds to groups of two or more PSUs in selection order, as they were selected systematically from an implicitly ordered list (see Lynn, 2009)
3321	UKHLS-GPS in Northern Ireland	Northern Ireland treated as a single stratum
3322 – 5117	UKHLS-EMB	Corresponds to the postal sectors in the high minority density domain as selections were made independently from each (see Berthoud et al, 2009)

Example using Stata

In Stata, to obtain estimates that correctly take into account the sample design the user need only specify the design variables using the svyset command, for example:

```
svyset a_psu [pweight=a_indpxus_xw], strata(a_strata)
```

Then, any compatible command simply needs to be prefixed with svy, for example:

```
svy: logistic depvar  variable1  variable2  variable3
```

Weighting adjustments for the Wave 2 release

A number of weights are provided for data users. They adjust for unequal selection probabilities, differential nonresponse and potential sampling error. A weighted analysis will adjust for the higher sampling fraction in Northern Ireland and for different probabilities of selection in the Ethnic Minority Boost sample, as well as for response rate differences between subgroups of the sample. Separate sets of weights are provided for the GPS and EMB sample component and the ex-BHPS sample. Considering the complexity of the study design, weights should be selected carefully, following the advice provided below. The weighting strategy is described in Lynn and Kaminska (2010).

The first part of this section covers the purpose of the weights and how to use the naming conventions for the weight variables to interpret and select the different weight variables from among a complex assortment. This is followed by the technical details of how weights were calculated.

If your aim is to generalise to the UK population, do not conduct unweighted analyses. For advanced users who want to model nonresponse in their own way, we provide design

weights (see below) which adjust the sample for unequal selection probabilities. Note that adjusting for the first wave nonresponse is different from adjusting for attrition and requires variables which have values for both responding households and never responding households.

In this release, we do not provide a weight for combining the former BHPS sample component with the General Population and Ethnic Minority Boost sample components. At this point, we recommend that analyses be carried out separately for the two samples. Note that the two samples represent slightly different populations as the BHPS does not represent people who were not resident in the UK in 1991 or their descendants.

Selecting the correct weight for your analysis

Given the complexity and multi-purpose nature of the UKHLS design we provide multiple weights to meet different needs of users. The weight for your analysis reflects the survey instrument, which is the source of the data being used in the analysis, and the analysis level (household or individual).

Each weight has been scaled to have a mean of one amongst cases eligible to receive the weight.

All weights follow a naming convention, designed to help users to pick the correct weight. The name of each weight reflects the wave for which the weight is calculated, level of analysis, data source and its nature (design weight, cross-sectional analysis weight or longitudinal analysis weight). The rules are described in the 'Naming Conventions for Weighting Variables' section below.

If your analysis uses only data from Wave 2, select the "xw" (cross-sectional) version of the weight. This weight is defined for all sample members who responded to the relevant survey instrument at Wave 2. If your analysis uses data from both Wave 1 and Wave 2, select the "lw" (longitudinal) version of the weight. This weight is defined for sample members who responded to the relevant survey instrument at both waves.

For individual level analysis you may want to combine information from different questionnaire sources. In this situation please select the weight suitable for the lowest level according to the hierarchy below:

Level of Analysis	Questions available for
4	household level (all enumerated individuals)
3	Adult proxy and main interview
2	Adult main interview only (no proxy)
1	Adult or youth self-completion interview

For example, if in one model from Wave 1 you use questions from the proxy and full interview as well as from the self-completion, then the correct weight will be a_indscus_xw –

the weight for the self-completion questionnaire as its level (1) is lower than the level for proxy and full interview (3).

Table 8. List of weight variables by analysis level, wave, and data source

Analysis level	Wave(s)	Data source	Analysis Weight
household	1	Household grid and/or household interview	a_hhdenus_xw
household	2	Household grid and/or household interview	b_hhdenus_xw
household	2	Household grid and/or household interview (BHPS)	b_hhdenbh_xw
individual	1	Household grid and/or household interview	a_psnenus_xw
individual	2	Household grid and/or household interview	b_psnenus_xw
individual	2	Household grid and/or household interview (BHPS)	b_psnenbh_xw
individual	1&2	Household grid and/or household interview	b_psnenus_lw
individual	1991-2008 &2	Household grid and/or household interview (BHPS, GB 1991)	b_psnen91_lw
individual	2001-2008 &2	Household grid and/or household interview (BHPS, UK 2001)	b_psnen01_lw
individual	1	Adult main and proxy interview	a_indpxus_xw
individual	2	Adult main and proxy interview	b_indpxus_xw
individual	2	Adult main and proxy interview (BHPS)	b_indpxbh_xw
individual	1&2	Adult main and proxy interview	b_indpxus_lw
individual	1	Adult main interview	a_indinus_xw
individual	2	Adult main interview	b_indinus_xw
individual	2	Adult main interview (BHPS)	b_indinbh_xw
individual	1&2	Adult main interview	b_indinus_lw
individual	1991-2008	Adult main interview (BHPS, Great	b_indin91_lw

	&2	Britain)	
individual	2001-2008 &2	Adult main interview (BHPS, UK)	b_indin01_lw
individual	1	Adult extra 5 minutes interview	a_ind5mus_xw
individual	2	Adult extra 5 minutes interview	b_ind5mus_lw
individual	1&2	Adult extra 5 minutes interview	b_ind5mus_lw
individual	1	Adult self-completion	a_indscus_xw
individual	2	Adult self-completion	b_indscus_xw
individual	2	Adult self-completion (BHPS)	b_indscbh_xw
individual	1&2	Adult self-completion	b_indscus_lw
individual	1	Youth self-completion	a_ythscus_xw
individual	2	Youth self-completion	b_ythscus_xw
individual	2	Youth self-completion (BHPS)	b_ythscbh_xw
For advanced users only			
household		(Wave 1 household design weight)	a_hhdenus_xd
individual		(Design weight)	a_psnenus_xd
individual		(Extra 5 minutes design weight)	a_ind5mus_xd
individual		BHPS inclusion weight for OSMs issued into UKHLS	b_psnenbh_li
individual		BHPS-2010 longitudinal enumerated person weight	b_psnenbh_lw

Not using weights

Note, that an unweighted analysis does not reflect the population structure correctly unless the assumptions below are true. It is suggested that researchers publishing or presenting unweighted estimates make these assumptions explicit.

If no weighting is used, an analysis of the UKHLS assumes:

that all estimates of interest are the same in Northern Ireland as in the rest of the UK; that people who live at an address with more than three dwellings or more than three households are the same as those who don't;

- that people who responded at Wave 1 are the same with respect to your estimates as those who did not; that people who continued to respond at Wave 2 are the same

as those who did not; and that people who responded to each particular instrument used in the analysis (individual interview, self-completion questionnaire etc) are the same as those who did not (see Lynn et al, 2012).

An unweighted analysis of the former-BHPS sample assumes:

that all estimates of interest are the same in each of England, Scotland, Wales and Northern Ireland;

- that people who live at an address with more than three dwellings or more than three households are the same as those who don't;
- that people who responded at Wave 2 of UKHLS in 2010 are the same with respect to your estimates as those who may have become non-respondents at any time since Wave 1 of BHPS in 1991.

We therefore strongly suggest conducting weighted analyses of the *Understanding Society* data.

Naming Conventions for Weighting Variables

The naming conventions for the will help users to select the weight they need or to interpret the purpose of some weight variables. The structure is as follows:

w_xxxyyzz_aa, where

w=wave

xxx=target population

yy=instrument

zzz=sample

aa=weight type

options for xxx:

hhd: household

psn: persons 0+

ind: persons 16+

yth: persons 10-15

options for yy:

en: enumeration (grid)

in: interview

px: interview or proxy

5m: “extra 5 minutes” items

sc: self-completion

ns: nurse visit

bd: blood

options for zz:

us: the GPS and Ethnic Minority Boost of the UKHLS sample

bh: BHPS sample

91: BHPS original sample starting in 1991 (England, Scotland and Wales)

01: BHPS sample starting in 2001 (original sample + Scotland and Wales boost + NI)

ip: Innovation Panel

Options for weight type aa:

lw: longitudinal analysis weight

xw: cross-sectional analysis weight

ld: longitudinal design weight

xd: cross-sectional design weight

li: longitudinal inclusion weight for BHPS sample. They represent the probability of a BHPS sample member being included in *Understanding Society* – i.e. equivalent to a design weight, but including non-response over 18 years or so. It is not an analysis weight.

Examples
a_indinus_xw is the cross-sectional analysis weight for individual interview data from Wave 1, representing the population of persons aged 16 or older.

b_indscus_lw is the longitudinal analysis weight for individual self-completion interviews from Wave 1 and Wave 2 representing the adult population who continuously lived in UK at the times of Wave 1 and 2.

Technical details
In this section we describe in turn how the weights were derived for:

- UKHLS GPS and EMB wave 1 weight;
- UKHLS GPS and EMB longitudinal weights;
- UKHLS GPS and EMB cross-sectional weights after Wave 1;

- BHPS longitudinal weights;
- BHPS cross-sectional weights.

UKHLS Wave 1 weights

The Wave 1 household level weights consist of two components: a design weight and nonresponse adjustment for household level nonresponse. Wave 1 individual level weights consist of four components: the design weight, nonresponse adjustment for household level nonresponse, individual level within-household nonresponse, and post-stratification to population characteristics. Each of the components is explained below.

Design weight

The design weight corrects for unequal probability of selection at a number of levels.

The household level design weight corrects for:

- Unequal selection probability due to the boost in Northern Ireland. The GPS selection probabilities in Northern Ireland are approximately twice those in other parts of the UK;
- Unequal selection probability due to the ethnic minority boost. Selection probabilities in the EMB part of the sample vary considerably between areas, depending on the estimated ethnic mix of the area and ethnic composition of the household. Additionally, households in high density areas with at least one ethnic minority member were weighted to account for combined probability of being selected as part of GPS or as part of EMB samples;
- The selection probability of households in a dwelling with more than 3 households or at an address with more than three dwellings is adjusted for the fact that only three such households were selected from the same address.

Individual level design weights correct for all the above with one specific difference: non-EM persons who live with EM persons in the same household have a chance to be selected only via the GPS part of the sample, and not via EMB. This means that non-EM persons in the EMB (who are TSMs) are given a design weight of 0 while non-EM persons in the GPS are given the household design weight. The weights for EM persons are adjusted for their dual probability of being part of GPS and EMB.

Individual level design weights for those eligible to answer the "Extra five minutes" is similar to the above design weight but differs in the following ways. It adjusts for the fact that the GPS-Comparison Sample is only $1/45^{th}$ of the GPS original sample; that all EM members in low-density areas were administered the "Extra five minutes"; and that EM members in high-density areas had a chance to be selected into either the GPS Comparison sample or the EMB. Similar to the above weight, non-EM persons were assumed to have a chance to be part of only the GPS Comparison Sample and not part of the EMB.

Household level Nonresponse Adjustment

Household level nonresponse adjustment is more complex than in other surveys given the large number of households which were selected as part of the EMB with unknown eligibility. Households who were selected as part of the EMB sample were screened on whether they contain at least one member of a relevant EM group (Berthoud et al., 2009). Given the low proportion of eligible households in EMB sample it is unrealistic to assume that all nonresponding households would be eligible, i.e. contain at least one EM member. To take

this into account we modeled eligibility and used this information in household nonresponse adjustments such that households which were more likely to be eligible had higher influence on the nonresponse correction. Note, that the predicted eligibility multiplied by the design weight is released for all the EMB sample households of unknown eligibility as part of a_hhdenus_xd. This will enable an advanced user to model Wave 1 household nonresponse taking into account the chance to be eligible among households of unknown eligibility.

To model eligibility we used predictors from the sampling frame and administrative neighbourhood data linked at a geographical level (for detailed description see below). After excluding ineligible addresses (like businesses or demolished and nonexistent addresses), the eligibility was modeled using only EMB households with known eligibility status (either screened out or screened in). This prediction was then extrapolated onto EMB households of unknown eligibility (e.g. not contacted). Given the limited number of selected addresses in Wales and Scotland and differences between countries in the available auxiliary variables (see below), we predicted eligibility using two models. The first included common predictors for England and Wales and eligibility was predicted for these two countries. The second was based on England, Wales and Scotland, using a more limited number of predictors. Eligibility was predicted for Scotland only from this model.

Following this, the probability of responding was estimated using backward stepwise logistic regression weighted by eligibility status (where the ineligible were excluded, those known to be eligible had a eligibility of one, and those with unknown eligibility had a weight proportional to the predicted probability of being eligible obtained from the above model). The predictors used in this model were the same as for the eligibility model and are described in detail below. Given that administrative neighbourhood data differs between England, Wales, Scotland and Northern Ireland, a separate model was implemented for each country. GPS and EMB response propensity was modeled together (which allowed us to model nonresponse within each country separately), but the indicator of EMB was retained in the model even if it was not statistically significant.

Predictors used for eligibility model and household level nonresponse correction come from the following sources:

- Sampling frame information, including such variables as sample month and geographical region;
- Predicted ethnic density of the postcode sector for five main ethnic groups in England, Scotland and Wales (as described in Berthoud et al., 2009);
- A wide range of indicators from Census 2001 and the most updated version of neighbourhood statistics as of summer 2011, linked separately for England, Wales, Scotland and Northern Ireland (see below).

The household nonresponse correction weight was calculated as the inverse of probability from the above model. This weight was multiplied by the household design weight to create the Wave 1 household level weight. The design effect was estimated using this weight showing that no truncation was necessary. The obtained weight was scaled to a mean of 1 and was named a_hhdenus_xw.

Neighbourhood statistics

For England and Wales the information was linked at Middle Layer Super Output Area (MSOA) or Lower Layer Super Output Area (LSOA) levels and was obtained from http://neighbourhood.statistics.gov.uk. The examples of linked information obtained from Census 2001 include the proportions in the MSOA of employed, retired, outright property owners, travellers to work using different types of transport, single household members, households with one car, people with different types of qualification and professional occupation, among others. Other linked information includes 2010 information on multiple deprivation indexes, on crime instances, 2009 information on inflow and net change of neighbourhood population, and the proportion of different allowance claimants, and 2008 information on hospital admissions and energy consumption.

For Scotland the information was linked at the data zone level from http://www.scrol.gov.uk/scrol/common/home.jsp and from http://www.scotland.gov.uk/Topics/Statistics/SIMD. From the Census 2001, information was obtained on population density, mean age, average household size and number of rooms per household in the data zone, as well as the proportions in the data zone born in Scotland and outside the EU, of different religious denominations, employed, unemployed and retired, disabled, those with different levels of qualification and types of occupation, and different types of accommodation, among others.

For Northern Ireland the information was linked at the Super Output Area (SOA) level and was obtained from http://www.ninis.nisra.gov.uk/. Examples of predictors obtained from Census 2001 at the SOA level include the average hours worked by residents, the average age of residents, percentages of residents with different level of qualifications, with different employment statuses, and with different types of marital status, among others. The predictors also include 2007-2009 information on multiple deprivation indexes.

Note, that using *Understanding Society* analysis weights (all but design weights), adjusts for household nonresponse bias in any estimate, to the extent it is related to the above mentioned variables.

Enumerated Individual Weight

The weight for analysis of enumerated individuals (a_psnenus.xw) is not equivalent to the household weight for all household members, as often happens in other household studies. This is because we have TSMs in Wave 1, who are not ethnic members selected into EMB part of the sample. Thus, the individual level design weight is not equal to the household level design weight for individuals in households containing a mix of EM and non-EM persons. The weight for the analysis of enumerated individuals is calculated as the product of individual level design weight a_psnenus_xd and the household level nonresponse correction (described above). The design effect was tested showing that no truncation was necessary. Weighted sample distributions were then compared to ONS mid-year estimates (with a correction for institutionalized population) and poststratification was implemented for the fully crossed matrix of gender by geographical region by 5-10 year age groups. Thus the individual level enumerated weight consists of:

The individual level design weight * the household nonresponse correction * the poststratification adjustment. The obtained weight is then scaled to have a mean of one.

Individual Level Nonresponse Adjustment

Five different individual level weights were prepared for users reflecting nonresponse occurring at different levels and different questionnaire instruments. Each individual level weight consists of:

The individual level design weight * the household nonresponse correction * the individual level nonresponse correction conditional on household response * the poststratification adjustment.

The individual nonresponse correction (conditional on household nonresponse) is modelled at three levels:

- For adult respondents (age 16 or older) who either completed the main interview or for whom a proxy interview was completed (for a_indpxus_xw);
- For adult respondents (age 16 or older) who completed the main interview only (for a_indinus_xw and a_ind5mus_xw);
- For respondents aged 10 or older who completed and returned the self-completion questionnaire (for a_indscus_xw and a_ythscus_xw).

Note, that the same model was used for respondents regardless of whether they were selected into GPS or EMB; that the response propensity is assumed to not depend on whether respondents received the "Extra five-minutes" or not; and that conditional on age (present in the model), the response to self-completion is assumed to have the same predictors for adults and youth (this assumption allowed modelling the response in each country separately, which wouldn't otherwise be possible for youth sample).

The individual level response, conditional on household response, was modeled using backward stepwise logistic regression separately for England, Wales, Scotland and Northern Ireland. The four models were implemented for each of the three levels described above. The predictors used in the models include all the predictors used for the household level nonresponse models and individual and household-level variables obtained from the household questionnaire, such as age and gender, marital and employment status, household size and presence of children in the household, as well as household expenditure on food and food outside, consideration of use of environmental energy, among others.

The individual-level non-response adjustment was obtained as the inverse of the predicted probability and was then multiplied by the relevant (either individual or "Extra-five minutes") design weight and by the household nonresponse correction. No truncation was deemed necessary as there were no extreme values substantially impacting design effects. The poststratification was implemented as described above in the individual level enumeration weight section, except that a greatly reduced matrix was used in the case of the "Extra-five-minutes" weight, due to the much smaller sample size for which this weight applies. After multiplying by the poststratification adjustment, each of the obtained weights was then scaled to a mean of one.

UKHLS longitudinal weights

Each of the five types of longitudinal weights (enumerated persons, proxy or main interview, main interview, self-completion, and "Extra-five minutes" interview) is based on the corresponding Wave 1 cross-sectional weight, with an additional adjustment for non-response at Wave 2. Each adjustment was based on a model of Wave 2 response conditional on Wave 1 response to the relevant instrument. For the enumerated persons

model, covariates were taken from the Wave 1 household grid and household questionnaire. In the model for proxy and main interviews, covariates were taken from the Wave 1 proxy interview (or the equivalent items from the main interview), household grid and household questionnaire. In both the model for main interviews and the model for adult self-completion questionnaires, covariates were taken from the Wave 1 main interview, household grid and household questionnaire. The adjustment weight was calculated as the reciprocal of the model-predicted response propensity. The Wave 1 weight had already adjusted for differential selection probabilities and differential Wave 1 response at both household and individual levels and, at the individual level, included a poststratification adjustment to mid-year population estimates by age, sex and region. The Wave 1 weight was then multiplied by the Wave 2 adjustment to create the Wave 2 longitudinal weight.

Newborns born to an OSM mother since the Wave 1 interview received the longitudinal enumerated person weight of their mother (reflecting the idea that the probability of observing the newborn is equal to the probability of observing the mother). The principle behind the longitudinal weights is that they are defined for each person who is observed at all of the relevant waves for which they were eligible. For this reason, newborns observed at Wave 2 receive a Wave 1 – Wave 2 longitudinal weight as they were enumerated at Wave 2, the only wave for which they were eligible.

UKHLS cross-sectional weights after Wave 1

The cross-sectional enumerated individual weights are based on the longitudinal enumerated individual weights, which are shared to temporary sample members (TSMs) and permanent sample members (PSMs) who entered the sample at Wave 2 through a weight-share method. Note, that only new TSMs and PSMs entering the study after Wave 1 receive a shared weight. TSMs who were present in Wave 1 (in the EMB sample) are given a cross-sectional weight of 0. This is done as the GPS part of the sample does not have an equivalent TSM group (OSM non-ethnic minority members living with TSM ethnic minority members). Giving a cross-sectional weight of 0 to Wave 1 TSMs maintains the balance of the whole sample.

These cross-sectional enumerated individual weights then serve as the base for the other cross-sectional individual-level weights, each of which (main, main or proxy, self-completion, youth) involves an additional adjustment for non-response to the relevant instrument conditional on enumeration[1]. The non-response models are therefore based on all eligible persons enumerated at Wave 2 (including TSMs and those OSMs who did not respond to the respective instrument at Wave 1), with covariates taken from responses to the UKHLS Wave 2 household grid and household questionnaire.

The cross-sectional weights for households (b_hhdenus_xw) are set equal to the minimum nonzero longitudinal enumerated person weight (b_psnenus_lw) amongst adults in the household, reflecting the idea that the probability of observing the household is equal to (or greater than) the probability of observing the person in the household who has the greatest probability of being observed.

[1] Note that there is no cross-sectional weight for the “Extra 5 minutes” questions as at Wave 2 these were only asked of sample members who had completed the main interview at Wave 1. Thus, the wave 2 longitudinal weight should be used for Wave 2 cross-sectional analysis.

BHPS longitudinal weights

Four weights will be continued from BHPS. Their variable names are changed. The corresponding weight variables are:

- Xlewght now called x_psnen91_lw,
- Xlewtuk1 now called x_psnen01_lw,
- Xlrght now called x_indin91_lw, and
- Xlrwtuk1 now called x_indin01_lw, where x represents the most recent UKHLS wave.

These weights are based on Wave 18 BHPS longitudinal weights, which account for the first wave household nonresponse, the first wave within household individual nonresponse (to enumeration or to an individual main questionnaire, respectively) and for individual nonresponse between the first wave and Wave 18 of BHPS. The base weights which reflect continuous enumeration (rlewght, a BHPS variable name) and continuous response to the main questionnaire (rlrght, a BHPS variable name) since 1991 are used for creating weights for longitudinal analysis starting 1991. Note, such an analysis excludes Northern Ireland as it was added to BHPS in 2001 and will also exclude the Scotland and Wales boost samples that were added in 1999. Similarly, the base weights which reflect continuous enumeration (rlewtuk1, a BHPS variable name) and continuous response to main questionnaire (rlrwtuk1, a BHPS variable name) since 2001 are used for creating weights for longitudinal analysis starting in 2001. Analysis using these weights will include all the BHPS samples. For more information on the BHPS weight calculation please refer to BHPS documentation (Taylor 2010).

For each of the Wave 18 weights an additional adjustment is applied to correct for attrition between Wave 18 of the BHPS and Wave 2 of *Understanding Society*, when the BHPS joined *Understanding Society.* The adjustment is the reverse of the estimated probabilities of participation (enumeration or response to main questionnaire) based on logistic regressions predicting participation at Wave 2 of UKHLS conditional on participation at Wave 18 of BHPS. The covariates used in the model predicting enumeration are from the BHPS Wave 18 household grid and household questionnaire. The same covariates plus covariates from the Wave 18 main questionnaire are used for predicting response to the UKHLS Wave 2 main questionnaire. Enumeration weights for newborn babies (biological, step or natural) born to an OSM mother since the time of the BHPS wave 18 interview are equal to their mother's enumeration weight. For "rising 16 year-olds" (OSMs who turned 16 between the time of the BHPS Wave 18 interview and the UKHLS Wave 2 interview and who could therefore be aged 16, 17, or even 18 at the time of UKHLS Wave 2), main response weights consist of the relevant longitudinal enumerated person weight, with an adjustment for the probability of main response at Wave 2 conditional on enumeration at Wave 2. The adjustment is the inverse of the response propensity predicted by a separate logistic regression model (based just upon all adults and inferred to rising 16 year-olds) using covariates from the Wave 2 household questionnaire and household grid. The base weight for rising 16-year olds correction is continuous enumeration since 1991 (b_psnen91_lw) for the BHPS 1991 main response weight (b_indin91_lw), and is the BHPS-2010 longitudinal enumerated person weight (b_psnenbh_lw – see next section below) for the BHPS 2001 main response weight (b_indin01_lw). The main response weight for each rising 16 year-old is then scaled by a constant factor so that the ratio of rising 16 year olds to older adults among main questionnaire respondents equals the equivalent proportion among all

enumerated respondents. The weights (x_psnen91_lw, x_psnen01_lw, x_indin91_lw and x_indin01_lw) are calculated by multiplying the respective BHPS Wave 18 weight and the adjustment, and are scaled to one.

BHPS cross-sectional weights

The BHPS cross-sectional weights are created as follows: we first model the chance of each BHPS OSM being issued into the UKHLS (reflected in b_psnenbh_li), then the chance of being in a responding household (complete the household grid and the household questionnaire) at Wave 2 of the UKHLS conditional on being issued (reflected in b_psnenbh_lw). The weight b_psnenbh_lw is then extrapolated to TSMs and PSMs through a weight-share method to create b_psnenbh_xw. The detailed procedure for creating these weights as well as cross-sectional individual response weights is described below.

The inclusion weight (b_psnenbh_li) was calculated separately for a) Northern Ireland and b) England, Scotland and Wales. For each, it has two components. For Northern Ireland, the first component consists of the BHPS Wave 11 cross-sectional weight, as this is the wave at which Northern Ireland first entered the BHPS. This component encompasses a design weight, poststratification and an adjustment for Wave 11 nonresponse. The second component is derived from a model of the propensity to be issued at UKHLS Wave 2 conditional on being enumerated in BHPS Wave 11. This therefore adjusts for all the stages of dropout between BHPS Wave 11 in 2001 and UKHLS Wave 2 in 2010. Model covariates were taken from the Wave 11 household grid and household questionnaire. This propensity was modelled as a single step from 2001 to 2010 because across-wave response patterns varied greatly between the sample members. There is no single BHPS wave since Wave 11 at which all the Northern Ireland sample members (of those issued to UKHLS) responded and therefore no other survey instrument that can provide model covariates for all relevant sample members. Similarly, for England, Scotland and Wales the first component consists of the BHPS Wave 9 longitudinal weight, as this is the wave at which the Scotland and Wales boost samples were added (so, all of the members of those samples who entered UKHLS were enumerated at that wave, as were the vast majority of members of the original BHPS Wave 1 sample who entered the UKHLS). This component therefore encompasses a design weight, Wave 1 poststratification and adjustments for non-response at each of the Waves 1 to 9 of the BHPS. The second component is derived from a model of the propensity to be issued at UKHLS Wave 2 conditional on being enumerated in BHPS Wave 9 and therefore adjusts for all the stages of dropout between BHPS Wave 9 in 1999 and UKHLS Wave 2 in 2010. Model covariates were taken from the Wave 9 household grid and household questionnaire.

BHPS OSM newborns since Wave 9 (England, Scotland or Wales) or Wave 11 (Northern Ireland) whose parents are both OSMs were then assigned a base weight equal to the smaller BHPS inclusion weight of their (OSM) parents in the child's 2010 (issued to UKHLS) household. This reflects the idea that the probability of the child entering the UKHLS sample equals the probability of at least one of his or her parents entering the sample, which in turn is equal to (or greater than) the probability of the parent who has the greatest probability of entering the sample. BHPS OSM newborns born to one OSM parent and one TSM parent were assigned a base weight equal to half of the OSM parent's weight in the child's 2010 (issued to UKHLS) household. The division by two reflects the idea that these newborns had double the chance of becoming BHPS OSMs, relative to people born to both OSM parents, as they would have been included had either their mother's or father's 1991 household been

sampled. For newborns who were observed with a single parent in a household in the first wave after their birth, the weight given was the parent's weight. This reflects a close to zero likelihood for the baby to be sampled via the other parent.

The adjustment for household non-response at UKHLS Wave 2 was derived from a model of enumeration at Wave 2 conditional on entering the UKHLS sample (i.e. being issued to the field for UKHLS Wave 2), in which covariates came from the Wave 9 household instruments for England, Scotland and Wales and the Wave 11 household instruments for Northern Ireland. The weight which reflects the chance of a BHPS OSM of being selected into the BHPS, to be issued into UKHLS and to be enumerated at Wave 2 of UKHLS is the BHPS-2010 longitudinal enumerated person weight (b_psnenbh_lw).

Finally, the BHPS cross-sectional enumeration weight (b_psnenbh_xw) was created through a weight-share method by sharing the BHPS-2010 longitudinal enumerated person weight to TSMs and PSMs.

The BHPS cross-sectional weights for main, proxy or telephone interview respondents (b_indpxbh_xw), main interview respondents (b_indinbh_xw) and self-completion respondents (adults (b_indscbh_xw) and youth (b_ythscbh_xw) each consist of the cross-sectional individual enumerated weight with an additional adjustment for non-response to the relevant instrument conditional on household response. These adjustments were based on logistic regression models with both individual-level and household-level covariates taken from responses to the UKHLS Wave 2 household grid and household questionnaire.

The BHPS cross-sectional household weight (b_hhdenbh_xw) is set equal to the minimum cross-sectional person enumerated weight (b_psnenbh_xw) amongst adults in the household.

Each weight has been scaled to have a mean of one amongst cases eligible to receive the weight.

Imputation of income variables

Understanding Society collects detailed information each wave on personal income. All individuals aged 16 or more are asked to report:

- wages,
- self-employment earnings,
- second job earnings,
- interest and dividends,
- pensions (National Insurance/state retirement pension, pension from a previous employer, pension from a spouse's previous employer, private pension/annuity, widow's or war widow's pension, widowed mother's allowance or widowed pension),
- benefits (severe disablement allowance, disability living allowance, war disablement pension, attendance allowance, carer's allowance, incapacity benefit, income support, job seeker's allowance, national insurance credits, child benefit, child tax credit, working tax credit, maternity allowance, housing benefit, council tax benefit, foster allowance/guardian allowance/rent rebate, rate rebate, employment and support allowance, respond to to work credit, sickness and accident insurance, in-work credit for lone parents and pension credit) and

- other income sources (educational grant, trade union and friendly society payment, maintenance or alimony, payments from a family member not living together, amount for rent from boarders or lodgers, rent from any other property).

These personal income variables can be summed to obtain the total personal income. Total household income can be computed from the personal total incomes of all household members.

Some of the income components can be missing. More precisely there can be three types of missing cases:

1. item non-response when individuals respond to the individual questionnaire but do not answer to some or all the questions on income components;
2. individual non-response when individuals fail to respond to the individual questionnaire;
3. household non-response when there is neither a household nor the individual questionnaire response.

For example at Wave 1 we have 59,466 individuals for whom at least the household questionnaire is available, and among these individuals 80.3% provided a personal interview, 5.5% have a proxy interview, whereas 14.2% had neither a proxy nor a personal interview. The item non-response rate for individuals who provided an individual questionnaire varies across income variables. It goes from a maximum of about 50% for self-employment earnings to zero for some of the benefit variables, and it is generally below 20% for the remaining income variables.

What do we impute?
In *Understanding Society* we do not impute income variables for non-responding households. Responding households are households for which the household questionnaire and information on the household composition/structure (household grid module) are available. We suggest that the user take account of household non-response via weighted estimates, described in Section 2 Weighting adjustments).

For individuals who respond to the individual questionnaire but do not provide answers to all income questions (item non-response), we impute the following personal income variables: wages, self-employment earnings, second job earnings, interests and dividends, pensions, benefits and other income sources.

For individuals for whom a proxy questionnaire is available, we impute total earnings and total income whenever missing. The proxy questionnaire is a short version of the individual questionnaire with questions on total earnings and total income as well as other variables.

Finally, for individuals in responding households for whom neither the personal nor the proxy questionnaire is available, we impute only the total personal income. This is not directly included in the data set, but is used in the imputation of total household income.

Based on these imputations we can compute total personal and household income for all individuals belonging to responding households.

For each income variable for which amounts are imputed there is a separate imputation flag variable (with a suffix _if instead of _dv indicating whether the variable is imputed. In most cases this takes the value 1 if imputed and 0 if not, but in the case of the following variables it shows the proportion of total income imputed: w_fimngrs_if, w_fibenothr_if, and w_fihhmngrs_if.

In the income data file there may be multiple receipts of income from the same source. For example, a respondent may have multiple pensions from a previous employer. These are imputed in a single step together, and the imputed values are in the variable w_frmnthimp_dv. This variable is set to inapplicable for the second and subsequent receipt of income from a single source.

Imputation procedures

The procedure used in *Understanding Society* is imputation by chained equations (ICE). Each income variable is imputed by stochastic regression imputation using as predictors a large set of auxiliary variables which includes income variables and other potential correlates such as personal and household socio-demographic characteristics. Some of these characteristics are missing and must also be imputed, but the released data contains imputed values only for the income variables.

Imputation by chained equations (ICE) allows for interdependence between income and auxiliary variables by considering univariate models estimated separately and sequentially (see Van Buuren et al 1999, and Ragunathan et al 2001). This method has been already used in some major household panel surveys such as the European Community Household Panel Survey.

In the following we describe the imputation by chained equations (ICE) adopted for item non-response in the individual (personal) and proxy questionnaires, and for individual non-response, that is, for those for whom there is neither an individual nor a proxy questionnaire available.

Item non-response on income variables in the individual questionnaire

The imputation of income variables in the individual questionnaire is performed considering a separate equation for each of the income components including each of the sources reported in the income data file.

We use log linear models for each of our income variables. The explanatory variables are a set of characteristics collected in the individual (personal) or household questionnaires. The specification of the models varies by income variable but it generally includes the following variables:

- personal socio-economic variables (age, sex, self-reported ethnic group, indicator for respondent born in the UK, marital status, education level, general health, current subjective financial situation);
- personal income variables (excluding the one used as the dependent variable);
- household characteristics (number of children in the household, house tenure, house type, household size);
- job characteristics (log number of hours normally worked per week, log number of hours per months in a second job, log years of job tenure, permanent or temporary job, occupation (soc 2000, 1 digit), number

employed at the current job workplace (for employees), number of employees if self-employed, whether is self-employed and hires employees, whether the employment organization is private or not (only for employees), type of ownership if self-employed (sole ownership or partnership), an indicator for whether annual business accounts are prepared for the Inland Revenue for tax purposes if self-employed);

- household variables reflecting economic situation (log amount spent on food from food shops in four weeks prior to interview, log amount spent on food eaten outside the home in four weeks prior to interview, log last year expenditure on domestic fuel (e.g. electricity and gas), number of bedrooms in the house, number of other rooms in the house, Council Tax band);
- government office regions.
- At Wave 2 we use the value for the same variable at Wave 1 (or Wave 18 of the BHPS) where this is available.

Furthermore, we use additional regression models to impute explanatory variables when missing. More specifically we use log linear regression for continuous variables and binary, ordered and multinomial logit models respectively for dummy, ordinal and unordered categorical variables. Finally, we consider interval regression when we have brackets rather than point information or when we have *a priori* information which allows us to bound the missing income variable. This is the case for dividends and interest, for which we have bracketed information.

The imputation by chained equations (ICE) starts by considering the following recursive (triangular) system of imputation equations,

where $Y_1,Y_2,\ldots,Y_k$ are the income and auxiliary variables to be imputed ordered from the one with the smallest percentage of missing values, Y_1, to the one with the largest percentage of missing values Y_k, X is a set of auxiliary variables observed for all individuals, α's and β's are parameters and $u_1,u_2,\ldots,u_k$ are random errors. Such a recursive system allows us to carry out the imputation separately for each variable and sequentially. The sequential procedure is given by the following steps:

1. estimation of the first equation and imputation of the missing values for Y_1,
2. estimation of the second equation using the imputed values to replace the missing values of Y_1, and imputation of Y_2,
3. repetition of estimation and imputation steps sequentially for each of the following equations until when all k variables, $Y_1,Y_2,\ldots,Y_k$ have been imputed.

We use stochastic imputation, that is, we draw the imputed values from the posterior predictive distribution of the variable to be imputed, conditional to the observed data. For more details about stochastic imputation we refer to Rubin (1987), Schafer (1997), and Kenward and Carpenter (2007).

This sequential estimation is consistent only if the recursive system is valid. Since this is not necessarily a valid assumption, ICE uses the imputed values produced using the above recursive system as starting values in an iterative imputation process. In other words, the starting values are used to begin a new cycle of imputations where each equation is estimated sequentially, but this time using as explanatory variables both X and all the imputed variables $Y_1,Y_2,\ldots,Y_k$ excluding the one used as dependent variable. At the end of this new cycle, a set of new imputed variables is produced and used to begin a further new cycle of imputations. These cycles of imputations are repeated until convergence. Notice that in practice some of the variables will be imputed by excluding some of the Xs and Ys variables because it does not always make sense to use all variables as predictors.

All variables are imputed as reported except for wages and self-employment income, where we convert amounts reported net to gross where gross is not reported, using a deterministic model based on the tax and national insurance system. Where net wages and self employment income are not reported, we convert from the gross amounts (reported or imputed) using the same model based on the tax and national insurance system. In computing total personal income it is assumed that all other sources are reported gross, or are not subject to taxation. We will in due course be producing net income estimates.

Item non-response for income variables in the proxy questionnaire
The only income variables reported in the proxy questionnaires are the total gross earnings and total gross income. We impute missing values for these two variables again using ICE. The imputation is based on the sample of persons responding to the individual questionnaire (where missing values have been replaced with the imputed values produced by ICE as explained in last section) together with the sample of individuals for whom a proxy questionnaire is available. The imputation process is comparable to the one described in last section. Since individuals answering the proxy questionnaires are asked to report income brackets rather than point values, we use interval regressions for both earning and income. We impute total gross earnings and total gross income using the explanatory variables described above.

Individual non-respondents with no proxy questionnaire
For individual non-respondents with no proxy questionnaire but in responding households, we use information from the household questionnaire to impute a total personal income. The procedure used is again the imputation by chained equations (ICE). We first impute the total gross income, then we impute the total net income using gross income as a predictor in addition to the other explanatory variables. The user should notice that the imputation of personal income for individuals for whom there is neither a personal nor a proxy questionnaire is based only on variables available in the household questionnaire. More precisely we use:

- individual socio-economic variables (age, sex, marital status, ethnicity, work)
- household socio-economic variables (household size, number of children in the household, whether there is nobody in the household who speaks English, whether the interview had to be translated, house type, an indicator for whether the person is owner of the house, the external condition of the address relative to the others, number of bedrooms in the house, number of other rooms in the house, value of the property for home owners, number of cars, number of durables, log last year's expenditure on domestic fuel (e.g. electricity and gas), amount spent on food eaten outside the home in four

weeks prior to interview, amount spent on food from food shops in four weeks prior to interview, weekly rent paid, whether the household can keep the accommodation warm enough)
- government office region, indicator for whether the area is a low density area for ethnic minorities.

Coding

Occupational coding for respondent's current or last occupations was carried out by NatCen. Coding was carried out on a case-by-case basis by trained coders, with 10% of the coding of SOC and SIC subject to a blind-coding check. Coding of parental occupations and respondent's first occupation was carried out within ISER, using the Computer Assisted Structured Coding Tool (CASCOT) system developed by Peter Elias. As a result of the six-figure codes attached via CASCOT, matching of the 1990 SOC coding with previous occupational classifications is now possible; in addition, special algorithms within CASCOT allow the re-coding of SOC codes into SEG, RGSC, Goldthorpe, Hope-Goldthorpe, Cambridge Scale and ILOISCO 88.

Several questions, e.g. country of birth, religion, political party, national identity, and citizenship had an "other, please specify" option. These responses were coded using an automated process.

Coding was also done for an open-ended question: "We've asked you a lot of questions but we also want to know what has happened in your **own** life that has been **especially important to you**. Can you please tell me anything that has happened to you, or your family, over the past year that has stood out as important?" The respondent could give up to four answers. The answers were recorded verbatim and manually coded for type of event and its subject.

3. File and variable information

The data release consists of multiple files in SPSS or Stata formats distributed by the UK Data Service. The list of files and their descriptors can be seen in the online documentation system (http://data.understandingsociety.org.uk/documentation/mainstage/dataset-documentation).

Information about the BHPS Sample Component

This release of Wave 2 data contains two *Understanding Society* samples:

1. The General Population and Ethnic Minority Boost sample for Waves 1 and 2.
2. The sample from the former British Household Panel Study (BHPS).

Both samples can be used for cross-sectional and longitudinal analyses. For both these purposes they will need to be analysed separately because of their different sampling histories. Separate weights are provided for the two samples as described in Section 2 Weighting adjustments).

The cases in the two samples can be distinguished using the variable b_memorig (for person level files) and b_hhorig (for household level files). These variables also allow the identification of different components of the BHPS sample (see below).

The questionnaires used for the two samples are the same. There are, however, a few differences in the data collected. One important issue is that the date of previous interview for GPS sample members who were interviewed at the previous was approximately 12 months earlier, while for the former BHPS sample the gap was between 13 and 27 months for sample members interviewed at Wave 18 of BHPS. This means that the reference period for history of events since the last interview will be longer for the BHPS sample.

For longitudinal analysis of the GPS sample, cases may be matched to Wave 1 data, available as part of this release from the UK Data Service, using the variable pidp, the *Understanding Society* cross wave person identifier.

However for the BHPS sample, a different identifier will need to be used: the variable pid, which is the BHPS cross-wave person identifier. The pid identifier is available in all person level files in the *Understanding Society* Wave 2 release and in the 18-wave BHPS longitudinal data set, available separately from UK Data Service (SN5151 British Household Panel Study, Waves 1-18, 1991-2009, http://www.esds.ac.uk/findingData/snDescription.asp?sn=5151).

While the great majority of BHPS sample cases who were interviewed in *Understanding Society* Wave 2 were previously interviewed at Wave 18 in 2008-9, there are a number who were last interviewed at an earlier wave. Information about the response status of BHPS sample members at each of the 18 waves is contained in the BHPS file XWAVEID.

The BHPS data set also contains a file called XWAVEDAT, which contains the values for stable variables (e.g. ethnic group, parent social class etc.). Because of some differences in variable definition, this information has not been copied across to the new *Understanding Society* file, also called XWAVEDAT. However in most cases values of these variables can be obtained by matching to the BHPS file. We hope to produce a harmonized version at subsequent release.

In matching to earlier waves of BHPS data it is important to be aware that variable names in the BHPS data set have slightly different formats:

- they are limited to eight characters
- there is no underscore separating the wave prefix from the main part of the name
- derived variables, imputation flags, weights and other special variables are not distinguished by _dv, _if suffixes.

However, most questionnaire variables which are carried in both surveys will have the same main variable name, though with a different wave prefix. Since the last wave of BHPS was Wave 18, the wave prefix is “R”. Thus if we wished to match Wave 2 work status (b_jbstat) on the file B_INDRESP to previous wave values, for the GPS sample we would match (using pidp) to A_INDRESP and use the variable a_jbstat, while for the BHPS sample we would match (using pid) to RINDRESP and use the variable rjbstat.

Variable information – overview basic and derived variables

Variable naming and labelling conventions

Most variables have a mnemonic name. Variables begin with a prefix designating the wave of data collection ("a_" for the first wave, b_ for the second wave). We have used "W_" to denote waves in general. We have attempted to keep the names of variables that came from the BHPS the same for the convenience of analysts, but this has not always been possible. Analysts should consult the BHPS documentation: https://www.iser.essex.ac.uk/bhps/documentation/volb/index.html.

Many derived variables are shown by the suffix '_dv'. Derived variables include variables copied over from one file to another for analytic convenience, variables that categorize a particular variable, e.g., age category; variables that combine information from multiple variables, e.g., body mass index from self reported height and weight. Information about how the derived variable is produced is shown in the notes for derived variables in the detailed variable view of the online documentation.

Additional codes denote different types of reasons for the lack of a valid response. These values have not been specified as missing in Stata or SPSS. However, these statistical packages have commands to assign values to missing for many variables simultaneously. Codes are

-9 Missing by error.

-8 Not applicable to the person or because of routing

-7 Proxy respondent. The question was not asked of proxy respondents or derived variable cannot be computed for proxy respondents.

-2 Refused

-1 Don't know

The meaning of other values is explained with the variable's value labels. There may also be notes in the detailed variable view of the online documentation system on the website (data.understandingsociety.org.uk/documentation/mainstage/dataset-documentation).

Learning about the study variables

There are multiple resources for learning about the study variables in order to plan analyses. These include the questionnaires and the module and variable views in the online documentation system.

Many of the basic (non-derived) variables can be learned about directly from the questionnaires. As was shown in Figure 2, the questionnaire has much useful information. Please note that in the questionnaire, the variable name does not have the wave prefix. It also shows the brief variable label, text of the question, source of the question and value labels. Showcards to help the respondent in answering are also marked as part of the questionnaire. You can go back and forth from the question view to the variable view.

Identifiers and useful variables

Households are identified by w_hidp, a wave specific variable with a different prefix for each wave. It can be used to link information about a household from different records within a wave, but cannot be used to link information across waves. Since the composition of households change between waves, the data do not include a longitudinal household identifier.

Individuals are identified by the personal identifier (pidp), which is consistent in all waves and can be used to link information about a person from different records belonging to one wave, or to link information from different waves. Individuals are also identified by w_pno – the person number within the household. The combination of w_hidp and w_pno is unique for each individual.

Table 9 lists some variables commonly used in analysis and may help the analyst to begin planning. Recall that the variables with the prefix w_ have the values for that wave. There is also the file XWAVEDAT which has variables with stable values. Variables in that file do not have the wave prefix. Analysts should also remember to consult the section on specifying the complex sampling variables from Section 2, Sample design variables and analysis and on weighting from Section 2, Weighting adjustments.

Table 9. Some useful variables

Variable	Description
w_hidp	Household identifier
w_hhsize	Household size
w_hsownd, w_tenure_dv	House owned or rented
w_hhtype_dv	Household type
w_fihhmngrs_dv	Gross household income in past 30 days
w_emboost	Ethnic minority boost flag
w_gpcomp	Ppopulation Sample Comparison with EM boost
w_hhresp_dv	Household response outcome
w_hhdenus_xw	Household cross-sectional weight
w_psnenus_xw	household grid and household interview
psu	Primary sampling unit
strata	Sampling strata
pidp	Cross-wave person identifier
w_country	Country or part of the UK

w_gor_dv	Government office region
w_ivfio_dv	individual response outcome
w_jbnssec8_dv	Social class (NS-SEC)
w_sex	Sex
w_dvage	Age
w_marstat	Legal marital status
mpid, fpid	cross wave identifier of natural mother, father
a_nchild_dv	number of natural children in household
w_jbstat	Current economic activity (employment status)
w_jbhas	Did paid work last week
w_ukborn	Born in the UK and UK country of birth
w_fenow	Still in further education
w_qfhigh	Highest educational qualification
a_health	Long standing illness or impairment
w_jbsoc00	Current occupation (SOC2000)

Documentation of derived variables

Derived variables are variables that are copied from one file to another for analytic convenience, or computed from one or more variables. Some are computed by the Blaise CAPI program to control the routing within the questionnaire. Others were computed for the purpose of analysts. Analysts should consult the description of derived variables that they plan to use in their analyses.

The derived variables are documented on the detailed variable view on the *Understanding* Society website. The documentation summarises the variables used in the computation of the derived variable. See the detailed view for a_scghq2_dv , a categorical or caseness expression of scores for the GHQ-12 as an example.

Proactive dependent interviewing was used in Wave 2 and will continue to be used in subsequent waves to increase efficiency of data collection and lessen respondent burden. Specifically, information reported at an earlier time is fed forward to the respondent to personalize the question. So rather than ask a question about current occupation with its complex probing by interviewers, the question might say, “the last time you were interviewed you said you were “specific occupation” are you still “specific occupation”? Feed-forward variables are used at both the household and individual levels. For example b_ff_hhsize feeds forward the household size from the previous wave (Wave 1). The variable b_ff_plbornc is the country of birth of the respondent fed forward from the previous wave.

Note the use of the prefix "ff". Some of the fed-forward variables were not used in the wording of a question but were used by the CAPI script to route respondents appropriately based on information from the previous wave.

Example code for matching files

We are including six examples of common data management tasks useful in analysing the data. Each task is illustrated with code for Stata. Because Stata is case-sensitive, we have not displayed file and variable names in upper case, but in lower case. Statements beginning with // are comments. The six tasks include:

- Distributing household level information to individual level
- Summarising individual level information at the household level
- Matching individuals within a household
- Using the egoalt file to create household composition variables
- Merging individual files across waves into long format
- Merging individual files across waves into wide format

Example 1: Distributing household level information to individual level

In this example we will distribute household level information to individuals in those households. We can do this by merging household level file (such as w_household) with an individual level file (such as w_indresp) within the same wave.

```
// open the household level file

use a_hidp a_hhsize using a_hhresp_ip, clear


// sort it on the household identifier, w_hidp

sort a_hidp


// save this temporary file

save hhinfo, replace


// open the individual level file

use pidp a_hidp  a_marstat using a_indresp_ip, clear


// sort it on the household identifier, w_hidp

sort a_hidp


// merge it with the earlier saved file on w_hidp. The output shows how many cases matched
```

```
merge m:1 a_hidp using hhinfo

// drop this variable – essential step

drop _merge

save final1, replace


// clean up unwanted files

erase hhinfo.dta
```

Example 2: Summarising individual level information at the household level
In this example we will summarise individual level information within a household (number of 18-24 year olds in the household) and then match that onto the household level file.

```
use a_hidp a_hhsize using a_hhresp_ip, clear

sort a_hidp

save hhinfo, replace

use pidp a_hidp a_dvage using a_indall_ip, clear


// create a variable that counts the number of 18-24year olds in each household

bysort a_hidp: egen n1824= sum(a_dvage>=18 & a_dvage<=24)


// keep only first observation for every household

bysort a_hidp: keep if _n==1


// keep only household level information

keep a_hidp n1824


// now merging this household information with the household level file

sort a_hidp

merge 1:1 a_hidp using hhinfo
```

```
drop _merge

save final2, replace


erase hhinfo.dta
```

Example 3: Matching individuals within a household
In this example we will match the information of wives onto that of their partners/spouses.

```
/* Open the dataset with information on all persons in responding households and keep only those persons who have a spouse/partner in the household*/

use a_hidp a_pno a_hgpart a_sex a_dvage using a_indall_ip if a_hgpart>0, clear

// rename the prefix a_ to something that would indicate that this information relates to the spouse or partner

renpfix a_ sp_

/* rename the spouse/partner pno variable to the respondent pno variable as this will be used to match on to the respondent information. Then sort and save the data*/

rename sp_hgpart a_pno

rename sp_hidp a_hidp

drop sp_pno

sort a_hidp a_pno

save spousepartner, replace


/* Again open the data with information on all persons in responding households*/

use a_hidp a_pno a_hgpart a_sex a_dvage using a_indall_ip if a_hgpart>0, clear

/* rename the prefix a_ to something that would indicate that this information relates to the respondent */

renpfix a_ r_

/* as we want to match on a_hidp and a_pno rename r_hidp and r_pno back to these */

rename r_hidp a_hidp

rename r_pno a_pno
```

```
// Now sort and merge with the spouse partner file

sort a_hidp a_pno

merge 1:1 a_hidp a_pno using spousepartner

drop _merge

save final3, replace

erase spousepartner.dta
```

Example 4: Using the egoalt file to create household composition variables
In this example we will create a variable that measures the number of siblings in the household using the egoalt file.

```
use b_hidp b_epno b_relationship using b_egoalt_ip, clear


// create a variable that counts the number of siblings in the household

bysort b_hidp b_epno: egen nsiblings = sum(b_relationship>=14 & b_relationship<=17)

lab var nsiblings "number of siblings in household"


// keep one observation per person

bysort b_hidp b_epno: keep if _n==1


sort b_hidp b_epno

save final4, replace
```

Now this information can be merged with any individual level file.

Example 5: Merging individual files across waves into long format
To match individual level files across *two* waves into a long format do the following (for more waves add wave specific prefix in the foreach statement):

```
foreach w in a b {

        // open the individual level file
```

```
        use pidp `w'_jbhas using `w'_indresp_ip, clear
        // drop the wave prefix from all variables
        renpfix `w'_
        // create a wave variable
        gen wave=strpos("ab", "`w'")

        // save one file for each wave
        save temp`w', replace
}

// open the file for the first wave (wave a_)
use tempa, clear
foreach w in b {
        // append the files for second wave onwards
        append using temp`w'
}

// save the long file
save final5, replace

// erase temporary files
foreach w in a b {
        erase temp`w'.dta
}
```

Example 6: Merging individual files across waves into wide format
To match individual level files across *two* waves into a wide format do the following (for more waves add wave specific prefix in the foreach statement):

```
use pidp a_jbhas using a_indresp_ip, clear

sort pidp

save temp, replace


foreach w in b {

        use pidp `w'_jbhas using `w'_indresp_ip, clear

        sort pidp

        merge 1:1 pidp using temp

        drop _merge

        sort pidp

        save temp, replace

}

save final6, replace

erase temp.dta
```

Preserving confidentiality

In preparing the data for the release we have taken steps to maintain the confidentiality of responses. These include not releasing the full date of birth and not releasing detailed geographic identifiers. Houshold income has been top coded. Open or narrative text, e.g., names of schools or employers, has not been released since it may indirectly identify individuals.

A Special License version of the data will be released through the UK Data Archive. The study has a Data Access Committee to take decisions on applications requesting access to electronic data and biological samples from *Understanding Society*. Its aim is to allow important research to proceed while minimising risks, particularly to Study participants.

Paradata

Some paradata, additional data collected about the interview process is available. These consist of call records, timings data and other information collected by the interviewers during the interview. The W_CALLREC data file has information on the number of calls made as well as the issue number, time and date and the outcome of each call. Information on the date of receipt of the case and the interviewer associated with each issue as well as the outcome at the end of each issue period is available in the file W_ISSUE. In addition to this, information collected in the address response form (ARF) by interviewers while contacting each household and asking household members to participate in the survey is available in W_HHSAMP. This includes data on the area surrounding the address, the type of accommodation and other information that the interviewer can observe about sampled addresses. Reasons for refusal are also available. Interviewers also collect some information

about the quality of the interview and persons present during the interview process. This is available along with substantive data collected during adult individual interviews (including proxy interviews) in W_INDRESP.

4. Data Access

We release the preceding waves of data when we make a new edition available. There have been some corrections to the earlier wave. The user should refer to the document made available by the UK Data Service, SN 6614 - Understanding Society: Wave 1, 2009-2010: Revision 1 (2012) for details.

We request that researchers using the data notify us about errors, inconsistencies, and other problems with the data identified during their use of the data. We make use of this information in improving the data. Please raise any issues relating to data or data analysis with our user support service; http://data.understandingsociety.org.uk/documentation/support. We will communicate information to members of the *Understanding Society* users group or via Frequently Asked Questions on the *Understanding Society* web page about data: http://data.understandingsociety.org.uk/.

The data are released through the UK Data Archive (UKDA) in SPSS and Stata formats. While documentation is released through the UKDA, we encourage users to consult the *Understanding Society* webpage. The documentation will develop over time. We plan to be developing specific guides about major content areas such as the biomeasures or cognitive measures, and guides for issues that are frequently problematic for users such as selection of appropriate weights.

Most of the Wave 1 has been released according to the conditions of the regular UKDA End User License: https://www.esds.ac.uk/aandp/access/licence.asp. A version of the Wave 1 to Wave 2 data has been released under conditions of the Special Licence (SL), SN 6931. Special License datasets are anonymised but contain more detailed information than End User Licence (EUL) data. The UKDA requires users to complete a set of forms with such detail as the intended use of the data. Researchers are asked to report publications resulting from the data.

Related *Understanding Society* releases are being prepared. One is a set of data products with information to link *Understanding Society* survey data with geographic units including Local Authority Districts, Area Classification for Output Areas, Travel to Work Areas, Westminister Parliamentary Constituencies, Rural-urban Indicators, Local Education Authorities, and Primary Care Trusts. For further information about these geographic units see (Office for National Statistics 2010) or the working paper on this topic (Rabe 2011). The geographical look-up tables are available under Special License conditions (http://www.esds.ac.uk/findingData/ukhlsSL.asp).

Notifications to ISER can be sent to info@understandingsociety.org.uk.

5. Citations and acknowledgements

Users should acknowledge both the UKDA and the Institute for Social and Economic Research in any publications arising from analysis of the data.

Citation of the data:

University of Essex. Institute for Social and Economic Research and National Centre for Social Research, *Understanding Society: Wave 1-2, 2009-2011* [computer file]. 4th *Edition*. Colchester, Essex: UK Data Archive [distributor], December 2012. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-4.

Citation of the User Manual:

McFall, Stephanie L. (ed.) (2012). *Understanding Society* –UK Household Longitudinal Study: Wave 1-2, 2009-2011, User Manual. Colchester: University of Essex .

People who participated in writing sections of the documentation included Jon Burton, Peter Lynn, Olena Kaminska, Gundi Knies, Randy Banks, Cheti Nicolletti, Laura Fumagalli, Jakob Petersen, and Nick Buck. Many people participated in preparing and processing the questionnaires and data. From the information technology side we recognize the contributions of Paul Groves, Paul Siddall, Geoffrey Angel, Tom Butler, Jeannette Chin, Elaine Prentice-Lane, and Catherine Yuen. From the survey research team, we recognize Noah Uhrig, Sarah Budd, and Emily Dix. A small group was active in contributing code for derived variables and flagging issues in using the data. They include Jakob Petersen, Cara Booker, Alexandra Skew, Mark Bryan, Mark Taylor, and Alita Nandi.

6. References

Berthoud, R., Fumagalli, L., Lynn, P., & Platt, L. (2009). *Design of the Understanding Society ethnic minority boost sample*, Understanding Society Working Paper No. 2009-02, Colchester: ISER, University of Essex. http://research.understandingsociety.org.uk/publications/workingpaper/2009-02.

Kenward, M., and J. Carpenter (2007), *Multiple imputation: current perspectives, Statistical Methods in Medical Research*, 16, 3, 199-218.

Lynn, P. (2009) Sample Design for *Understanding Society*. *Understanding Society Working Paper 2009-01*. Colchester: University of Essex. http://research.understandingsociety.org.uk/publications/working-paper/2009-01.pdf .

Lynn, P. Burton, J., Kaminska, O., Knies, G., and Nandi, A. (2012) An initial look at non-response and attrition. *Understanding Society Working Paper 2012-02.* Colchester: University of Essex. http://research.understandingsociety.org.uk/publications/working-paper/2012-02.

Lynn, P. and O. Kaminska. (2010) Weighting strategy for *Understanding Society. Understanding Society Working Paper 2010-05.* Colchester: University of Essex. http://research.understandingsociety.org.uk/publications/working-paper/2010-05.

Office of National Statistics (2010) 'Midyear population estimates 2009'. June 24, 2010 Edition, http://www.statistics.gov.uk/statbase/product.asp?vlnk=15106.

Rabe, B. (2011) Geographic identifiers in Understanding Society. Understanding Society Working Paper 2011-01. Colcheser: University of Essex. http://research.understandingsociety.org.uk/publications/working-paper/2011-01.

Ragunathan, E, T., Lepkowski, J. M., van Hoewyk, J., and Solemberger P. (2001), A Multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, 27, 1, pp.85-95.

Rubin, D. B. (1987): Multiple imputation for nonresponse in surveys. New York: Wiley.

Schafer, J. (1997): Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

Taylor, M. F. (ed) (2010) British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. Colchester: Universtiy of Essex.

van Buuren, S., H. C. Boshuizen, and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681–694.